Analyse des accidents de la route pendant l'année 2021 en fonction du lieu, de la période, de l'age et du sexe

Groupe FLAL

Contents

Sources de nos jeux de données : https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2021/

Alex Delagrange, Léo Bouvier, Lucas Giry, Farah Seifeddine

!!!!! A faire !!!!! :
- Télécharger le dossier “data” dans le même répertoire que les fichiers FLAL_stats.rmd et data.rmd
- Run le fichier data.rmd
- Knit le fichier FLAL_stats.rmd en PDF

Contexte : En France en 2021, le début d’année est accompagné d’un confinement, il est interéssant de voir quels sont les chiffres à propos des accidents mais aussi, pour une étude plus globale de prévention, savoir quels sont les facteurs qui influencent la gravité des accidents.

Le jeu de données vient du site data.gouv.fr.

Ainsi, pour progresser de manière structurée dans notre étude, nous nous sommes posés la question suivante :

Quels sont les facteurs importants impactant la gravité des accidents en France en 2021 ?

Après avoir établi les différents facteurs nous avons effectué plusieurs croisements pour expliquer la gravité : Les usagers (age et sexe), les vehicules (catégorie), le lieu (régions et départements), l’environnement (équipement de sécurité et état de la route)

Quelques éléments nécessaires à la compréhension :

Dans notre étude, un accident corporel est défini comme un accident survenu sur une voie ouverte à la circulation publique, impliquant au moins un véhicule et ayant fait au moins une victime ayant nécessité des soins. Cette définition prend en compte tous les accidents qui ont des conséquences physiques pour les personnes impliquées, qu’il s’agisse de blessures mineures ou de décès, ce qui nous permettra d’analyser les facteurs de risque des accidents de la route.

Nous étudions les accidents de la route en France en 2021 et un individu de notre jeu de données est une personne touchée par un accidents.

Nettoyage et représentation du jeu de données

En ce qui concerne la représentation des données nous avons choisi de représenter toutes les données dans une grande dataframe dont on récupérera les colonnes necessaires pour les différents facteurs.

Pour nettoyer notre jeu de données, nous avons d’abord examiné les valeurs manquantes, les doublons et les valeurs aberrantes. Par la suite nous avons enlevé les lignes de la dataframe sans données dans lesquelles les colonnes étaient remplies de -1. Nous avons aussi créer une nouvelle variable “region” en fonction de la valeur de la variable “dep”, qui correspond à un code départemental. Après ce processus, nous avons constaté qu’environ 0,04 % des lignes ont été éliminées en raison de valeurs manquantes ou de doublons.

En plus, 50 % des colonnes ont été éliminées car elles ne contenaient pas des informations pertinentes ou utiles pour notre analyse. Par exemple, la variable occutc qui indique le nombre d’occupants dans le transport en commun s’est avérée inutile, ainsi que le type de moteur du véhicule. Pour finir, il fallait enlever les caractères spéciaux des régions.

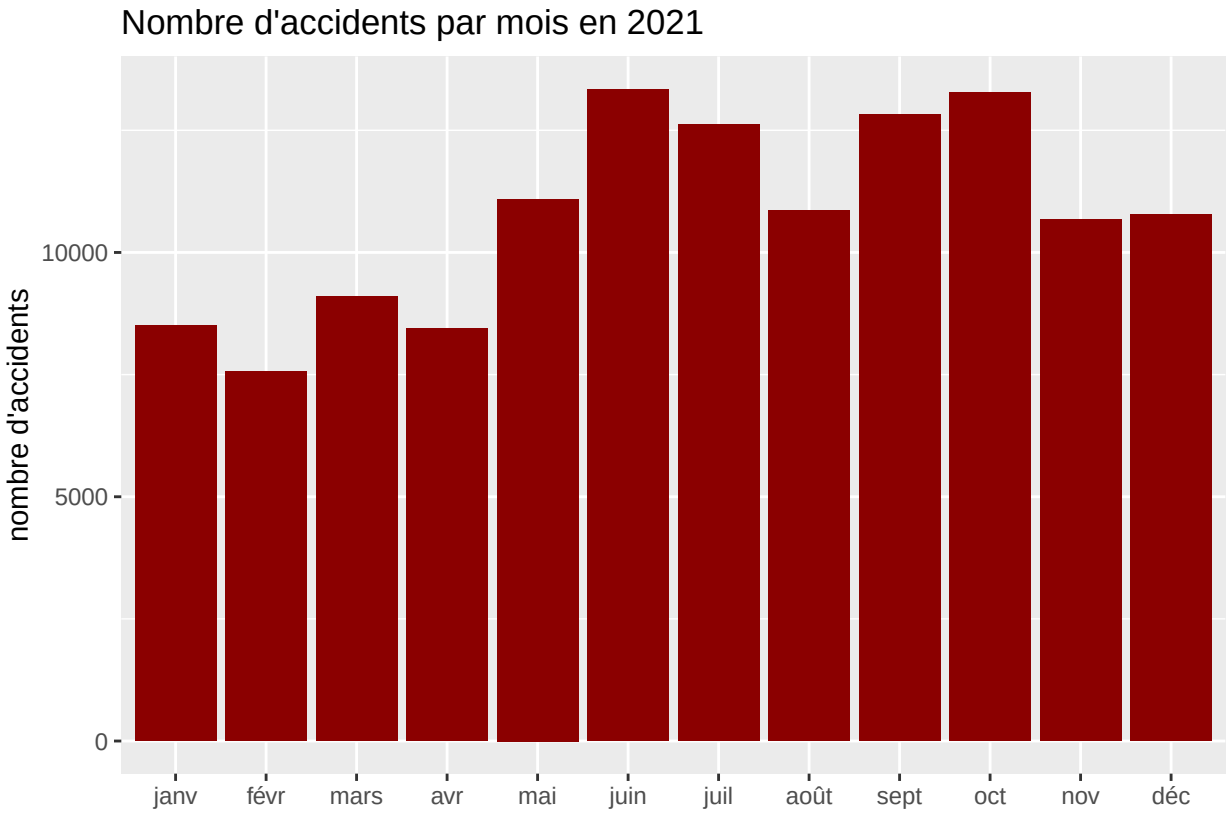
Ce nettoyage nous a permis d’avoir un jeu de données plus cohérent, propre et exploitable pour notre analyse ultérieure dont voici un aperçu :

```
## # A tibble: 6 x 27
##   Num_Acc     id_ve~1 num_veh place  catu  grav  sexe an_nais trajet secu1 secu2
##   <chr>       <chr>   <chr>   <dbl> <dbl> <dbl> <dbl>   <dbl>  <dbl> <dbl> <dbl>
## 1 2021000000~ 201Â 7~ B01         1     1     3     1    2000      1     0     9
## 2 2021000000~ 201Â 7~ A01         1     1     1     1    1978      1     1    -1
## 3 2021000000~ 201Â 7~ A01         1     1     4     1    1983      0     1    -1
## 4 2021000000~ 201Â 7~ B01         1     1     3     1    1993      0     1    -1
```

```
## 5 2021000000~ 201Â 7~ A01          1     1     1     1    1995      1     1     0
## 6 2021000000~ 201Â 7~ A01         10     3     3     2    1959      4     0    -1
## # ... with 16 more variables: secu3 <dbl>, catv <dbl>, obsm <dbl>, jour <dbl>,
## #   mois <dbl>, hrmn <time>, lum <dbl>, dep <chr>, com <chr>, agg <dbl>,
## #   col <dbl>, lat <dbl>, long <dbl>, vma <dbl>, age <dbl>, region <chr>, and
## #   abbreviated variable name 1: id_vehicule
```

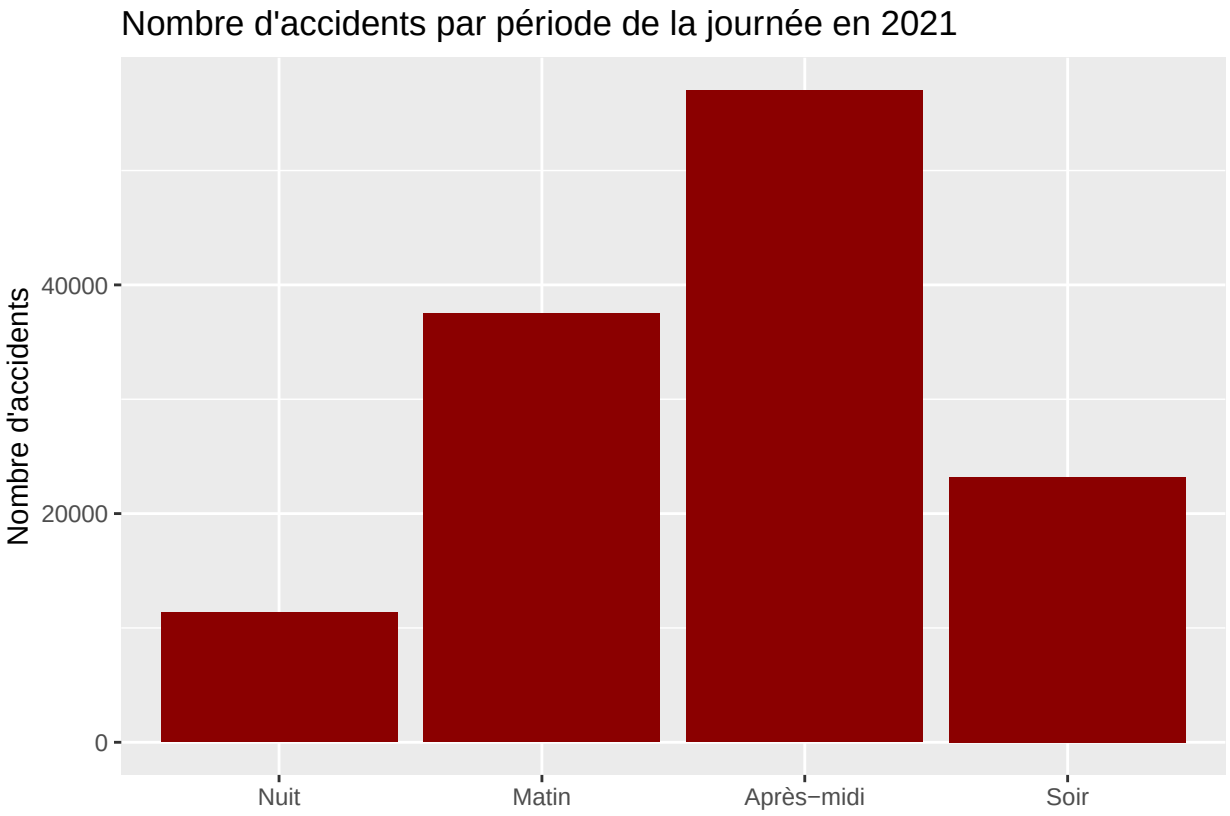
Etude globale sur les accidents en 2021 en France

On regarde d'abord juste une étude globale sur le nombre d'accidents en France en fonction du mois de l'année.



Ici on voit un gros pique en juin qui pourrait s'expliquer par la fin du confinement de 2021 ce qui a poussé les gens à sortir et donc à plus prendre la voiture donc forcément il y plus d'accidents.

On peut aussi regarder le nombre d'accidents en fonction de la période de la journée : nuit, matin, après-midi ou soir

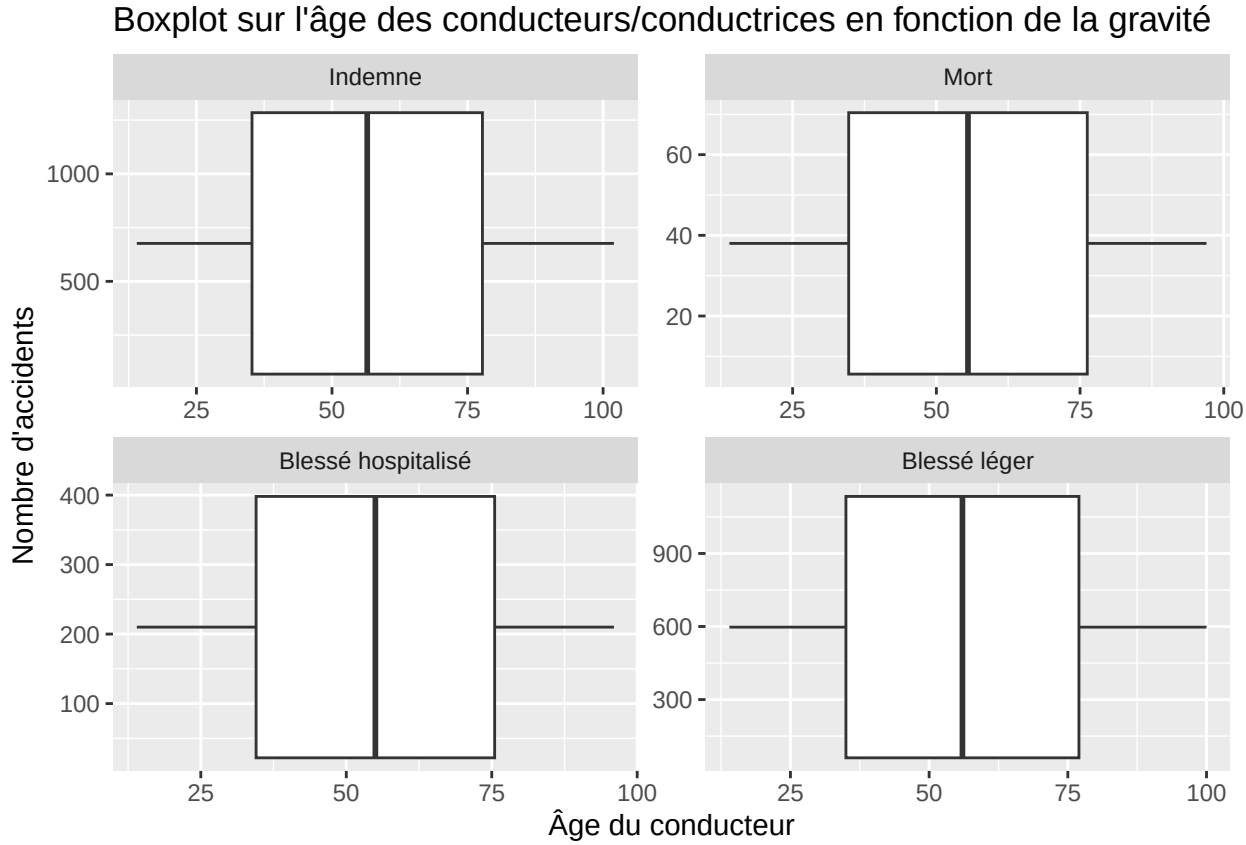


On s'attend à voir ce graphique comme résultat, avec plus d'accidents en après-midi et le matin puisque c'est dans ces moments là de la journée que l'activité française est au plus haut avec les départs au travail et les enfants à l'école.

Est ce que les usagers influencent la gravité des accidents ?

Commençons par étudier si l'âge du conducteur est en corrélation avec la gravité de l'accident

```
## `summarise()` has grouped output by 'grav'. You can override using the
## `.groups` argument.
```

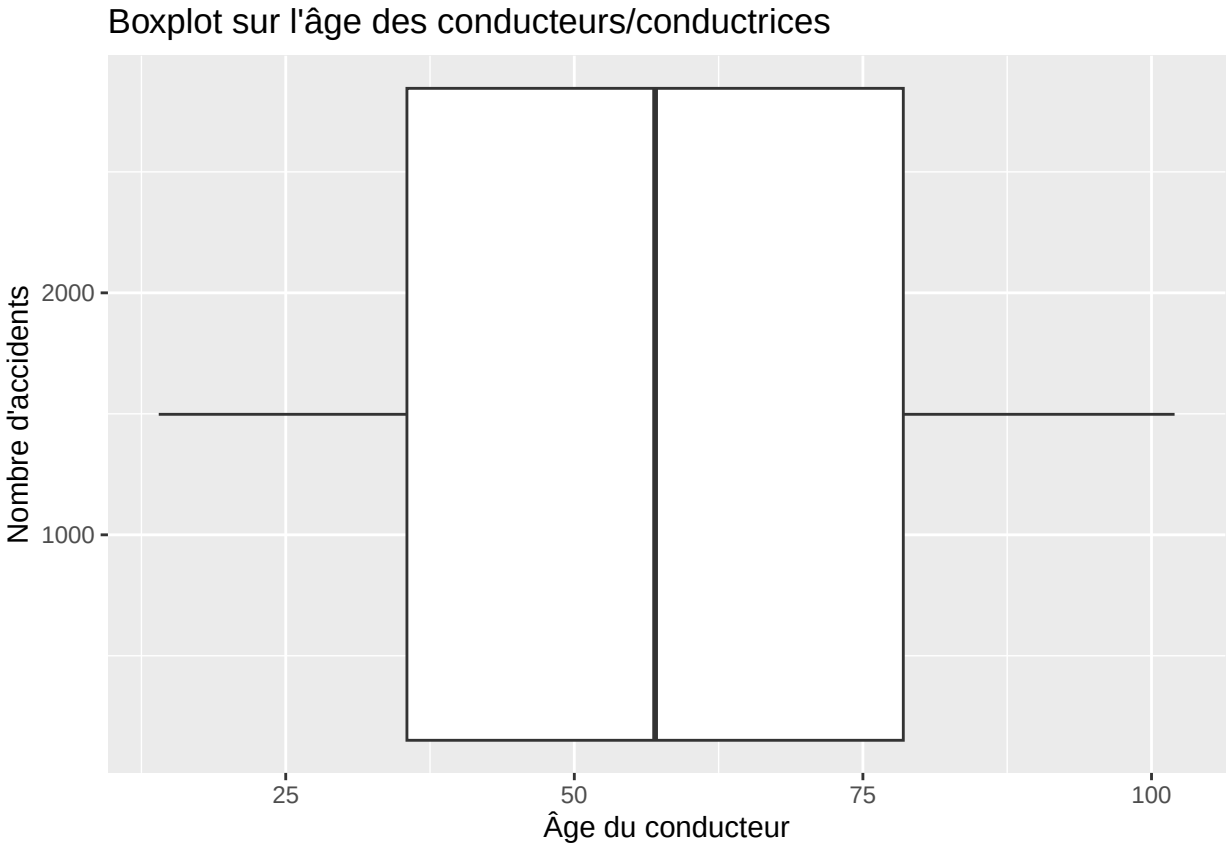



```
## sub_df$grav: Indemne
## $stats
## [1]  14.0  35.0  56.5  78.0 102.0
##
## $n
## [1] 86
##
## $conf
## [1] 49.17384 63.82616
##
## $out
## numeric(0)
##
## ------------------------------------------------------------
## sub_df$grav: Mort
## $stats
## [1] 14.0 34.5 55.5 76.5 97.0
##
## $n
## [1] 84
##
## $conf
## [1] 48.25953 62.74047
##
## $out
## numeric(0)
```

```
##
## ------------------------------------------------------------
## sub_df$grav: Blessé hospitalisé
## $stats
## [1] 14.0 34.5 55.0 75.5 96.0
##
## $n
## [1] 83
##
## $conf
## [1] 47.88947 62.11053
##
## $out
## numeric(0)
##
## ------------------------------------------------------------
## sub_df$grav: Blessé léger
## $stats
## [1]  14  35  56  77 100
##
## $n
## [1] 85
##
## $conf
## [1] 48.80225 63.19775
##
## $out
## numeric(0)
```

- Le nombre d'observations varie de 83 à 86.
- La médiane de l'âge est d'environ 35 ans pour tous les groupes de gravité.
- Le groupe "Indemne" a la valeur minimale la plus élevée et le groupe "Blessé léger" a la valeur maximale la plus élevée. - Cela indique que les conducteurs indemnes ont tendance à être plus âgés que les autres groupes de gravité, tandis que les conducteurs blessés légèrement ont tendance à être plus jeunes.
- Il n'y a pas de valeurs aberrantes pour chaque groupe de gravité.

Voici ce que nous pouvons observer sur l'ensemble du jeu de données :

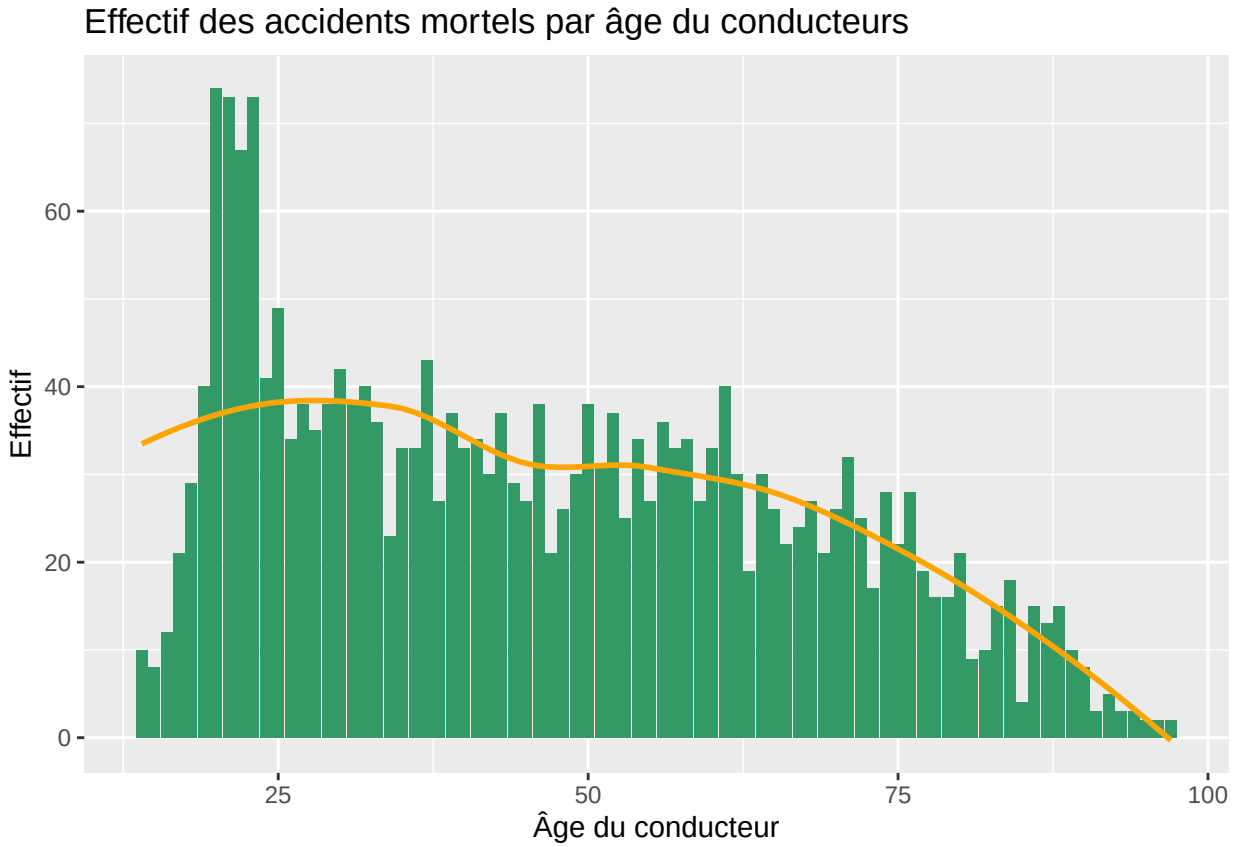


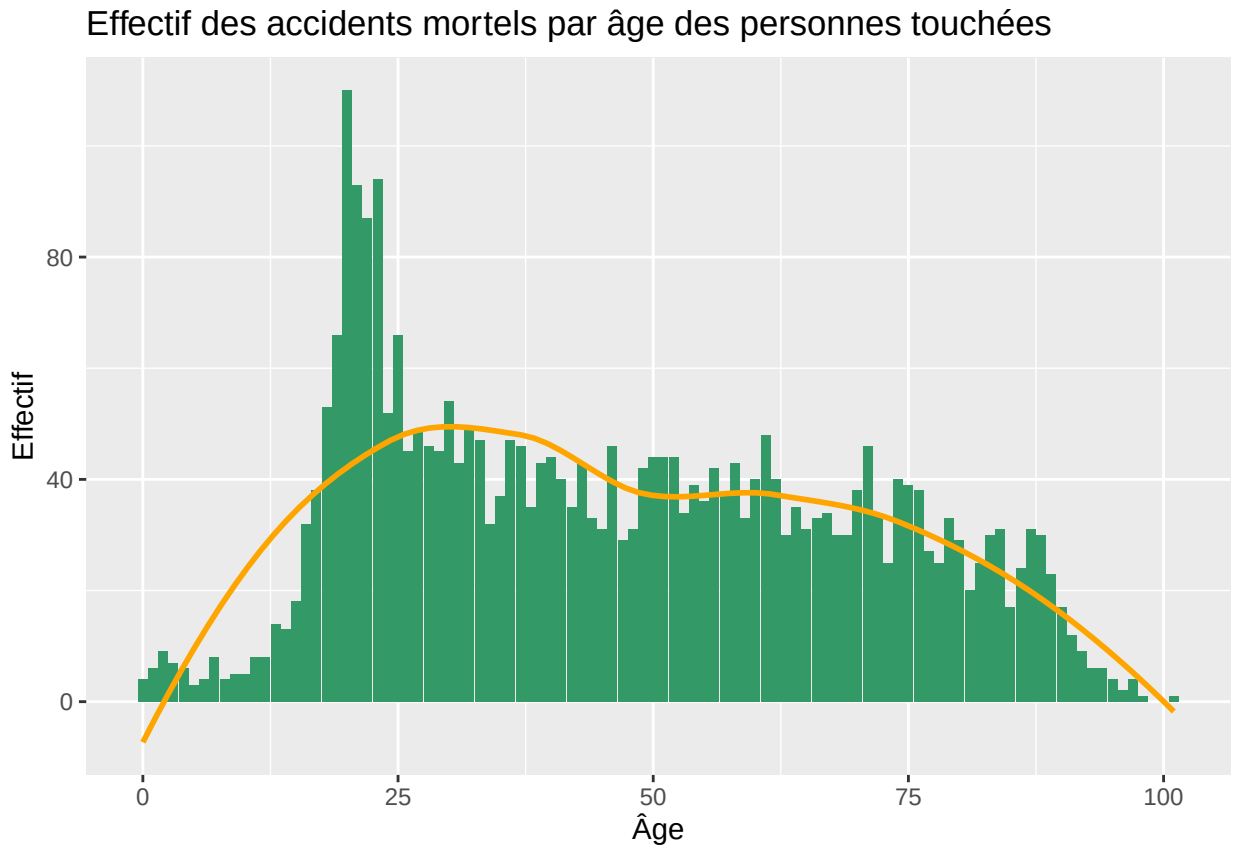

```
## $stats
## [1]  14.0  35.5  57.0  78.5 102.0
##
## $n
## [1] 87
##
## $conf
## [1] 49.71607 64.28393
##
## $out
## numeric(0)
```

Sur 126086 accidents de la route en 2021, 50 % des conducteurs ont moins de 57 ans. 50 % d'entre eux ont entre 35.5 et 78.5 ans.

Dans ce cas, l'intervalle de confiance est [49.7,64.3], ce qui signifie que l'on peut être raisonnablement sûr que la moyenne de l'âge dans la population dont l'échantillon a été prélevé se trouve dans cette plage avec une probabilité de 95%.

Pour un autre point de vue sur le nombre d'accidents par age (conducteur ou non conducteur) :

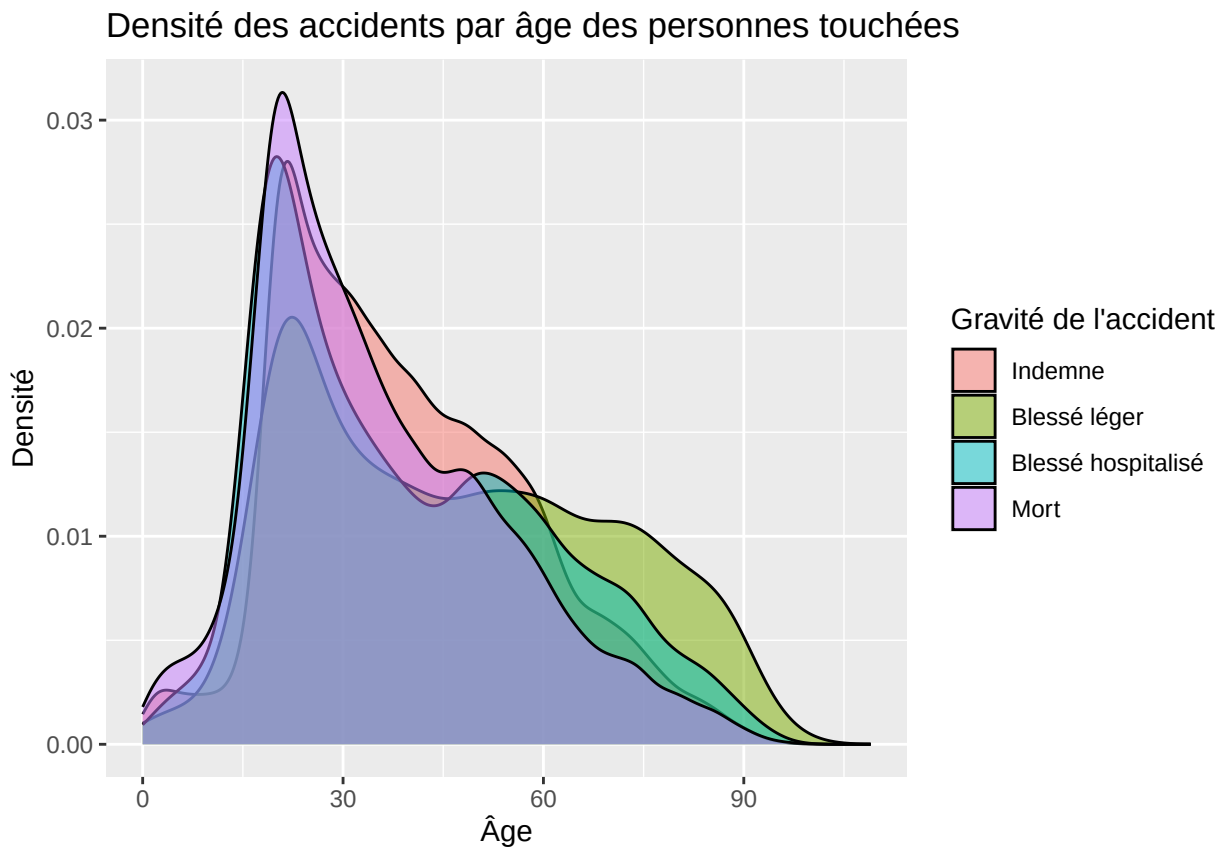

Effectif des accidents mortels par âge du conducteurs
Effectif
60
40
20
0
25
50
75
100
Âge du conducteur



On constate la même croissance du nombre d'accidents mortels (conducteur ou non) de 0 à 23 ans, les personnes touchées sont plus nombreuses, ce qui est normal il y a les passagers ajoutés. La tendance décroît fortement de 23 à 40 ans avant de rester plus ou moins au même niveau jusqu'à environ 70 ans. Pour les conducteurs, on constate une décroissance nette plus tôt que chez les personnes touchées, une hypothèse serait que plus l'âge augmente, moins il y a de conducteurs.

Le fort pique autour de 20 ans pourrait s'expliquer par l'âge de l'obtention du permis de conduire qui se traduit par un manque d'expérience en tant que conducteur. Cela pourrait aussi s'expliquer, dans le cas des personnes touchés par le fait que pas 100% des jeunes d'environ 20 ans ont une voiture, donc une personne de 20 ans avec une voiture va plus avoir tendance à amener avec lui ses amis donc un accident d'une voiture d'une personne de 20 ans compte possiblement 2, 3 voire 4 personnes.

De ces données nous pouvons en déduire le graphique de densité suivant :



De manière générale, le graphique de densité montre comment les distributions de l'âge des conducteurs diffèrent selon la gravité de l'accident. On remarque un fort pic de densité autour de 20 ans ce qui laisse penser à une forte corrélation entre l'age et la gravité de l'accident.

Pour déterminer si il y a une corrélation ou non entre l'âge et la gravité de l'accident, nous avons établi une régression logistique que voici :

```
##
## Call:
## glm(formula = grav_bin ~ age, family = binomial(), data = sub_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4426  -1.3180   0.9822   1.0296   1.2298
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.6048416  0.0129956   46.54   <2e-16 ***
## age         -0.0066719  0.0003048  -21.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 170975  on 126085  degrees of freedom
## Residual deviance: 170495  on 126084  degrees of freedom
## AIC: 170499
```

```
##
## Number of Fisher Scoring iterations: 4
```

Les résultats de la régression logistique indiquent que l'âge est significativement associé à la gravité des accidents. Plus précisément, pour une unité d'augmentation de l'âge, la log-odds d'avoir une gravité plus élevée diminue de 0,00667. Cela peut être interprété comme une diminution de la probabilité d'avoir une gravité plus élevée pour chaque année supplémentaire.

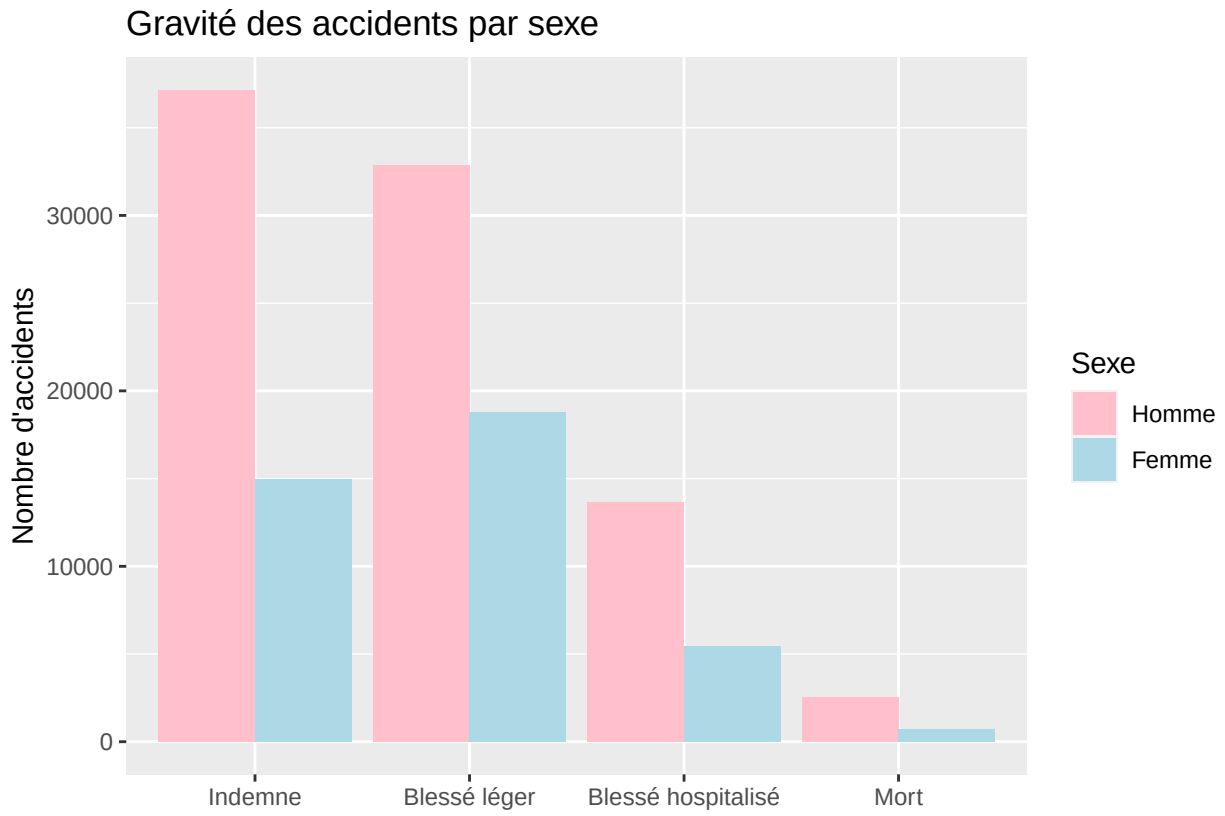
Le rapport des deviances (null deviance et residual deviance) montre que le modèle ajuste bien les données, car il y a une réduction significative de la deviance résiduelle par rapport à la deviance nulle. En outre, l'AIC est relativement faible, ce qui indique que le modèle est un ajustement approprié pour les données.

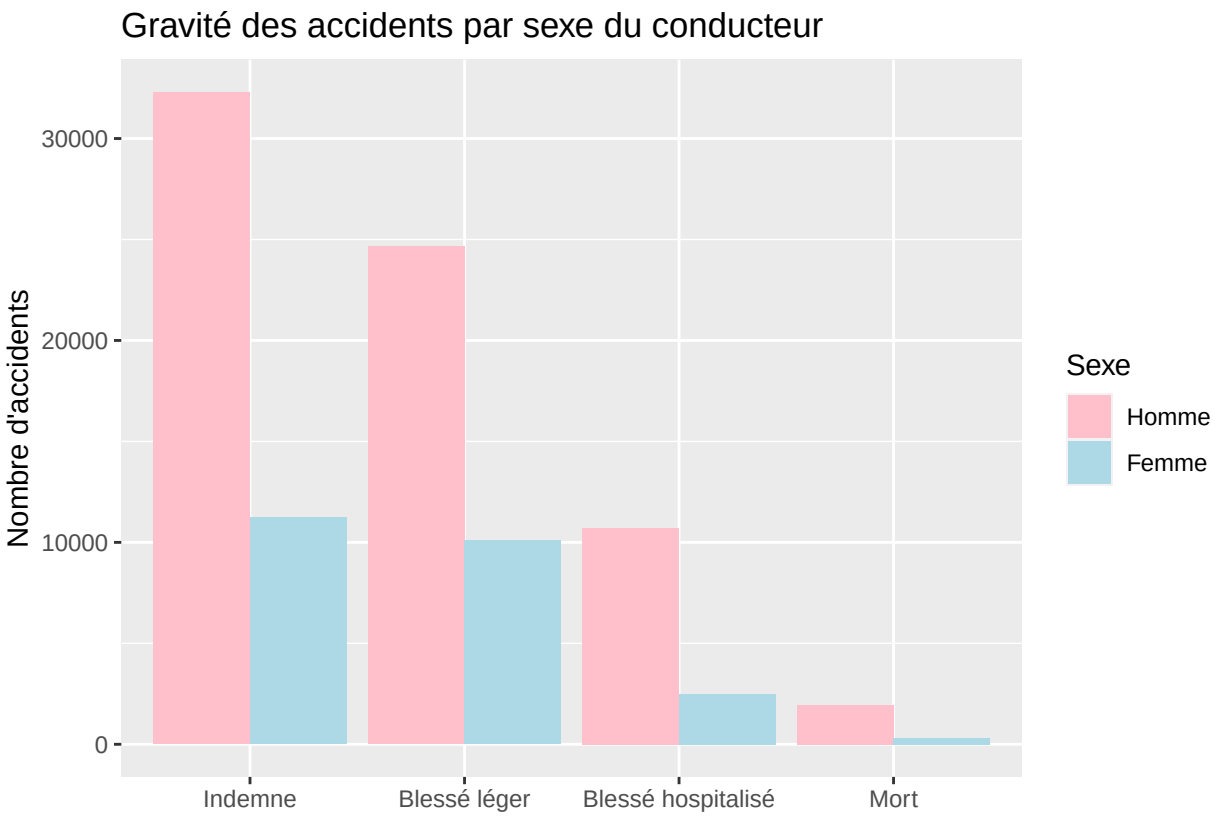
Le test de significativité indique que la relation entre l'âge et la gravité de l'accident est très significative (p-value $<$ 2e-16), ce qui renforce l'idée que l'âge est un facteur important à prendre en compte dans l'évaluation de la gravité des accidents de la route.

En résumé, les résultats de la régression logistique soutiennent l'idée que l'âge est significativement associé à la gravité des accidents de la route et que le modèle est un ajustement approprié pour les données.

Est-ce que le sexe influence la gravité des accidents ?

Ici, à l'aide de représentations graphique et de test du Chi2 nous allons observer si le sexe et la gravité ont une corrélation ou non en distinguant les personnes touchées et les conducteurs/trices.





On remarque que, conducteur.trice ou non, la tendance reste la même au niveau de la gravité. La quantité d'accidents reste plus importante chez les hommes que chez les femmes. Cependant, nous n'avons pas la quantité totale de conducteur.trice en France donc on ne peut pas avoir une proportion du nombre d'accidents.

```
##
##  Pearson's Chi-squared test
##
## data:  table(df_conducteurs$sexe, df_conducteurs$grav)
## X-squared = 686.69, df = 6, p-value < 2.2e-16
```

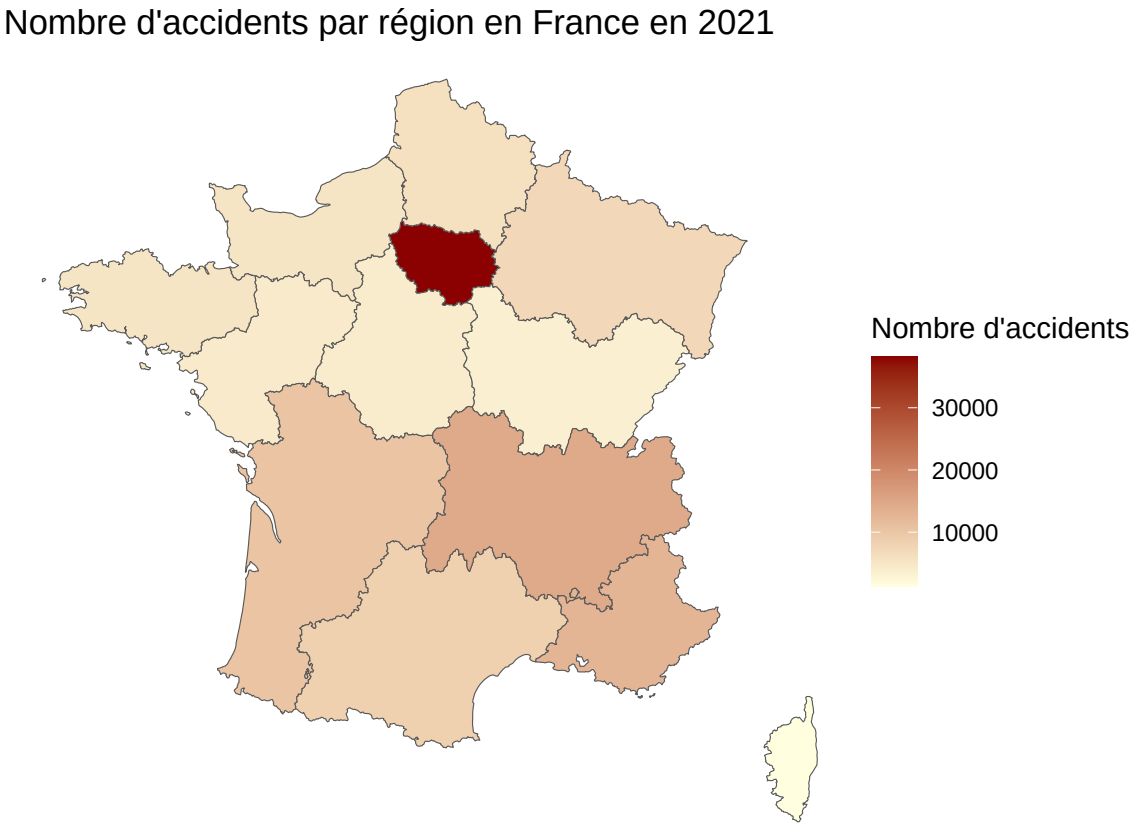
Puisque la p-value est inférieure au seuil de significativité communément utilisé de 0,05, on peut rejeter l'hypothèse nulle selon laquelle il n'y a pas d'association entre le sexe et la gravité des accidents de la route. On peut donc conclure qu'il y a une association significative entre le sexe et la gravité des accidents de la route en France en 2021.

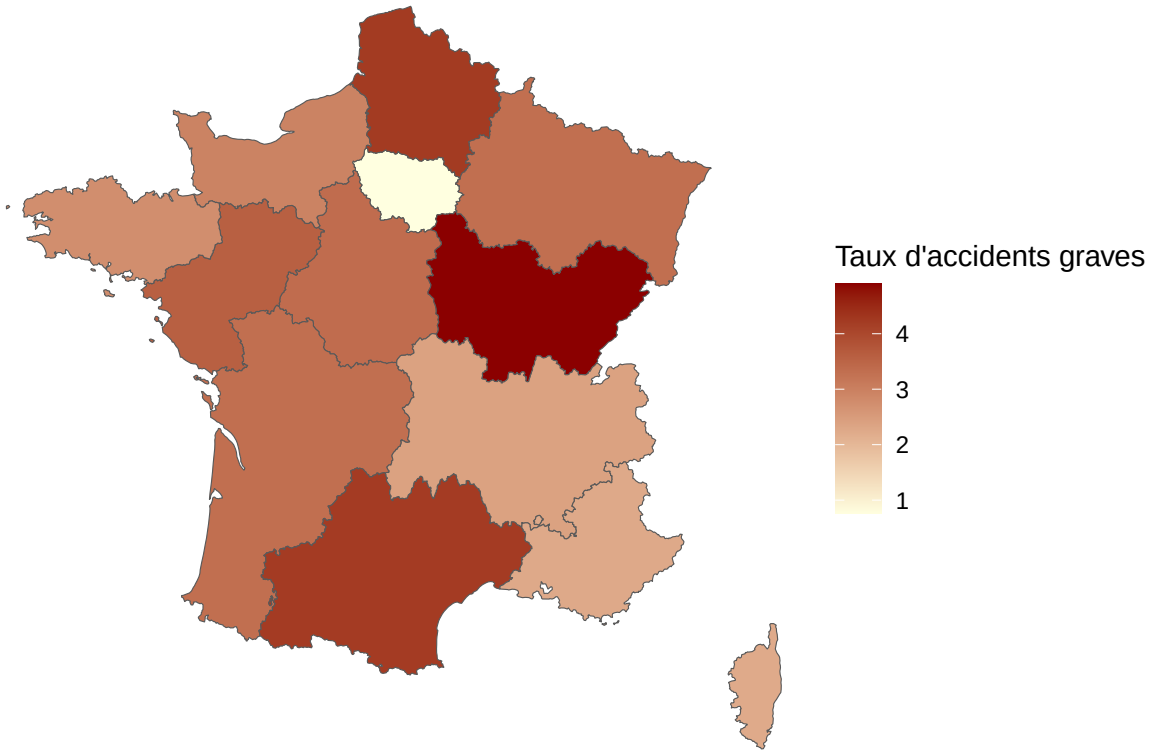
Cela signifie que le sexe des conducteurs est statistiquement significatif pour prédire la gravité des accidents de la route en France en 2021. Cependant, il est important de noter que les résultats de ce test ne permettent pas de déterminer la direction de l'association (c'est-à-dire, si les accidents graves sont plus fréquents chez les hommes ou chez les femmes).

Est ce que le lieu influence la gravité des accidents ?

Un autre facteur de gravité pourrait être le lieu de l'accidents, voyons si il y a des lieux où la gravité des accidents est plus forte.

```
## Reading layer ‘regions-version-simplifiee’ from data source
##   ‘https://raw.githubusercontent.com/gregoiredavid/france-geojson/master/regions-version-simplifiee.g
##   using driver ‘GeoJSON’
## Simple feature collection with 13 features and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -5.103601 ymin: 41.36705 xmax: 9.559721 ymax: 51.0884
## Geodetic CRS:  WGS 84
```






```
##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 6065.5, df = 39, p-value < 2.2e-16
```

Le résultat du test montre une statistique de test de 6065.5 et un degré de liberté de 39, ce qui donne une p-value inférieure à 2.2e-16.

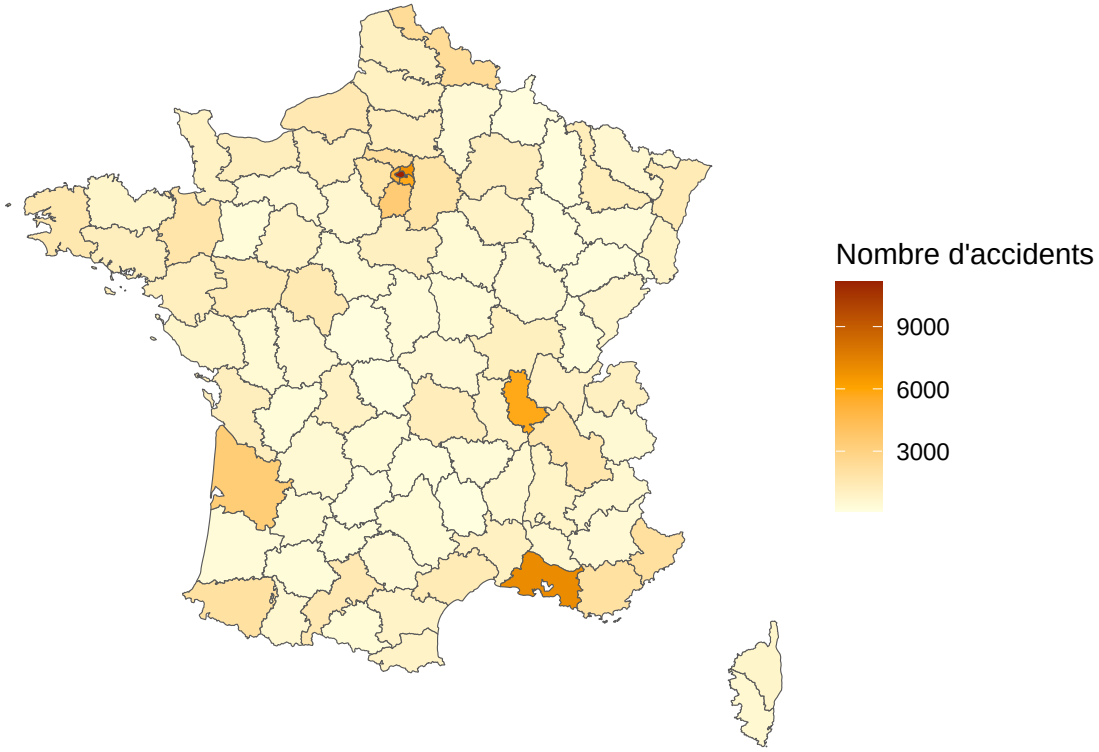
La p-value est très faible, ce qui suggère qu'il y a une forte association entre la région et la gravité de l'accident. Autrement dit, la gravité des accidents semble varier significativement selon la région où ils se produisent.

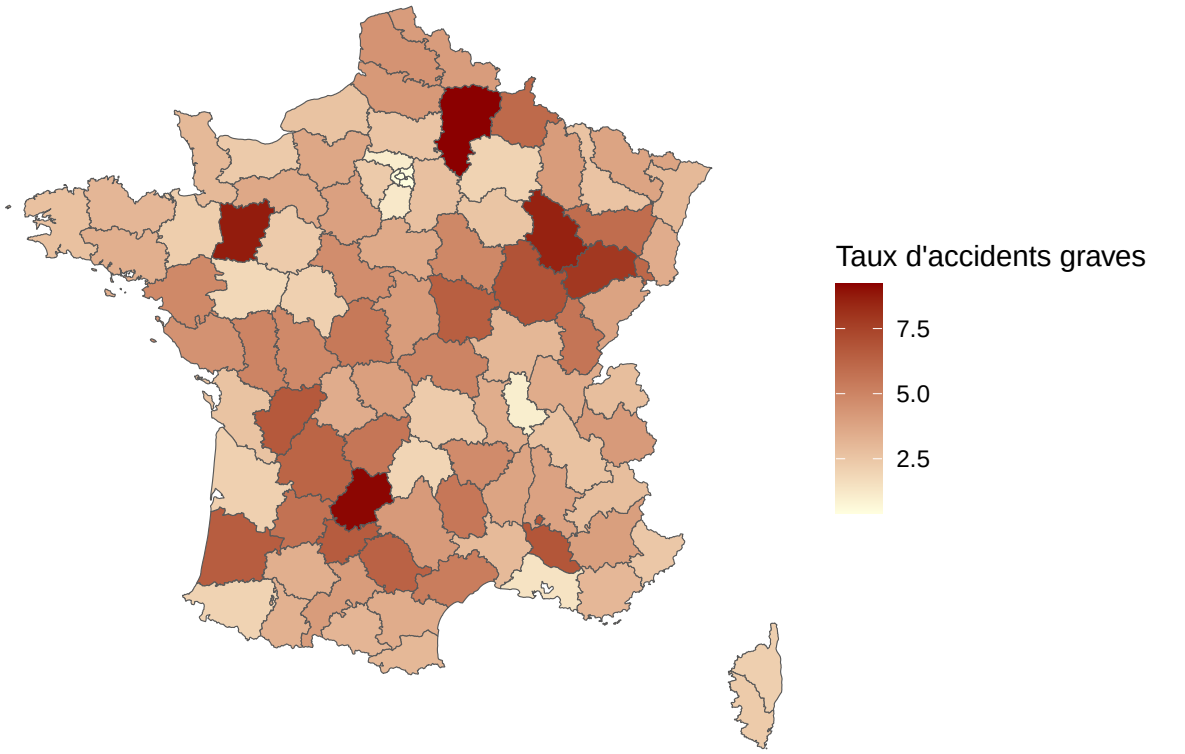
La région Parisienne a le plus de nombre d'accidents, mais c'est aussi la plus peuplée, on constate à la suite que cette région a le plus d'accidents mais les moins graves. On peut émettre l'hypothèse que la limitation de vitesse peut peut-être influer la gravité des accidents.

Seulement, notre variable sur la limitation de vitesse n'est pas utilisable en vue des valeurs au-delà de 130 km/h.

```
## Reading layer 'departements' from data source
##   'https://france-geojson.gregoiredavid.fr/repo/departements.geojson'
##   using driver 'GeoJSON'
## Simple feature collection with 96 features and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -5.138001 ymin: 41.36216 xmax: 9.559226 ymax: 51.08854
## Geodetic CRS:  WGS 84
```

Nombre d'accidents par département en France en 2021



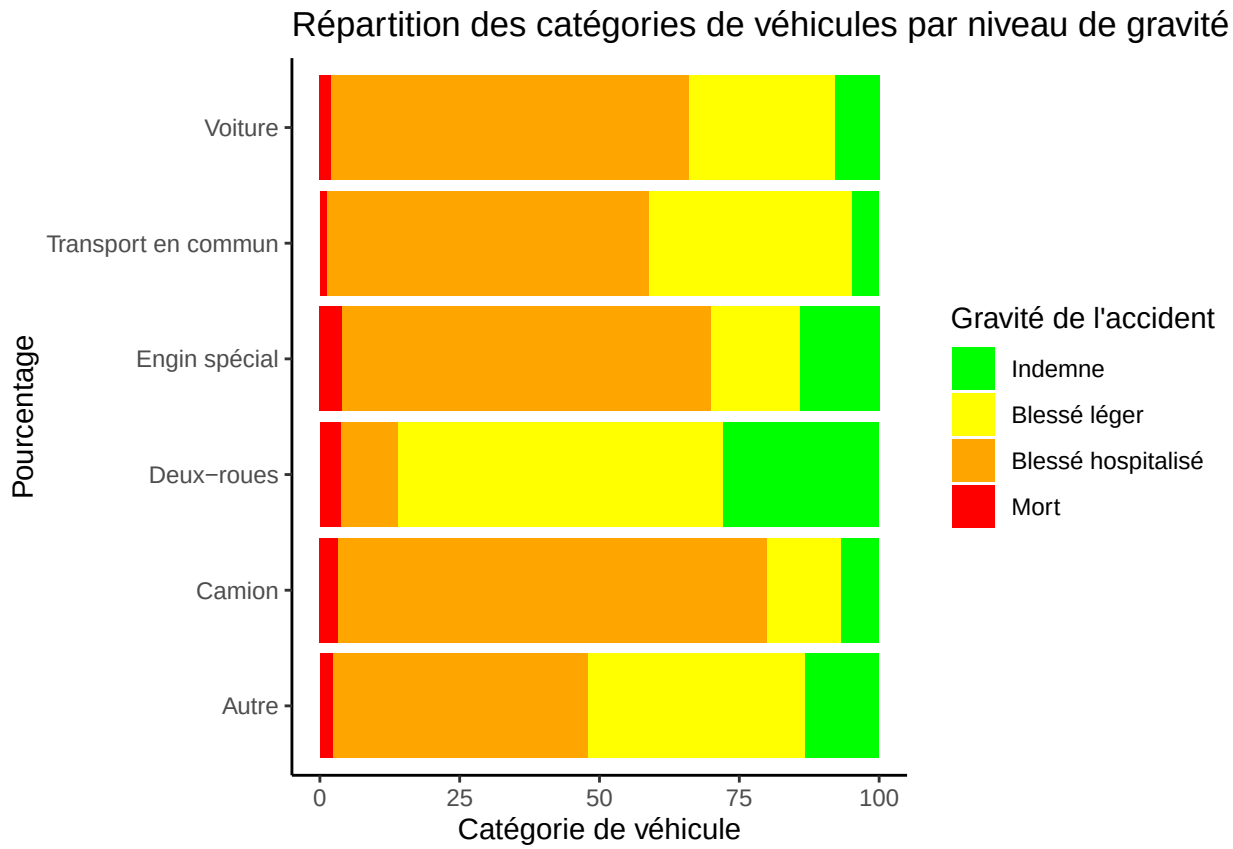



```
##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 10960, df = 318, p-value < 2.2e-16
```

Dans les deux résultats, la statistique de test est très élevée (6065.5 pour les régions et 10960 pour les départements), ce qui suggère qu'il existe un lien significatif entre la région/departement et la gravité de l'accident. On remarque que les deux cartes précédentes sont quasiment inverse c'est-à-dire que dans les zones où le nombre d'accidents est élévé, le taux d'accidents mortels l'est moins. On remarque ceci notamment pour la région d'île de france et pour le département du rhône. Cela pourrait s'expliquer par le fait que dans les grandes villes, les accidents se font à des vitesses réduites.

Est ce que l'environnement influence la gravité des accidents ?

Est ce que la catégorie de véhicule influence la gravité ?




```
##
## Call:
## glm(formula = grav_binary ~ categorie, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8727  -0.5807  -0.5807  -0.4624   2.3695
##
## Coefficients:
##                               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                  -1.694609   0.008903 -190.339  < 2e-16 ***
## categorieCamion              -0.487282   0.060194   -8.095 5.72e-16 ***
## categorieDeux-roues           0.925468   0.017959   51.533  < 2e-16 ***
## categorieEngin spécial        0.184850   0.125430    1.474    0.141
## categorieTransport en commun -1.050340   0.106775   -9.837  < 2e-16 ***
## categorieVoiture             -0.513298   0.036831  -13.937  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 118832  on 129092  degrees of freedom
```

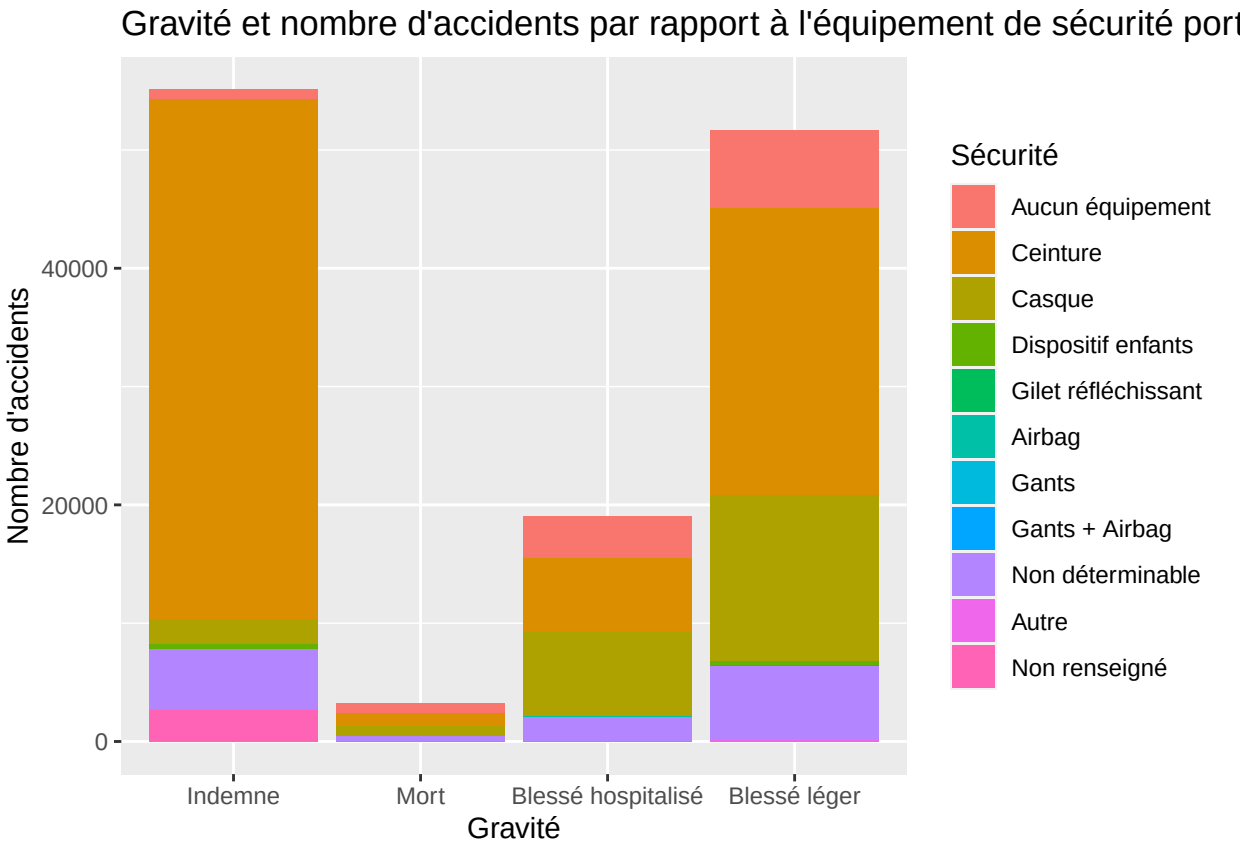
```
## Residual deviance: 115606  on 129087  degrees of freedom
## AIC: 115618
##
## Number of Fisher Scoring iterations: 5
```

La première partie de la sortie donne quelques informations sur la qualité de l'ajustement du modèle, notamment les résidus de deviance. Dans l'ensemble, les résidus semblent assez faibles, ce qui suggère que le modèle s'adapte bien aux données.

La deuxième partie de la sortie présente les coefficients estimés pour chaque niveau de la variable explicative (les différentes catégories de véhicules). Les coefficients indiquent l'effet de chaque niveau de la variable explicative sur la probabilité de gravité de l'accident. Par exemple, la variable categorieTransport en commun a un coefficient négatif (-1.050340), ce qui signifie qu'être impliqué dans un accident avec un transport en commun diminue la probabilité de gravité de l'accident par rapport aux autres catégories de véhicules.

En somme, ces résultats indiquent que la catégorie de véhicule est un facteur significatif pour prédire la gravité de l'accident, et que certains types de véhicules ont une probabilité de gravité plus élevée ou plus faible que d'autres.

Est ce que l'équipement de sécurité influence la gravité ?



On constate que, parmis les accidents de la route de notre jeu de données, une grande majorité a un équipement de sécurité équipé tel que la ceinture ou un casque pour les 2 roues.

```
##
##                     Aucun équipement      Ceinture       Casque
```

```
##   Indemne                1.417217101 79.834143863  3.812513610
##   Mort                  25.815470643 34.731283007 25.722273998
##   Blessé hospitalisé    18.889179984 32.963059995 36.772334294
##   Blessé léger          12.797739982 46.982450030 27.085389215
##
##                      Dispositif enfants Gilet réfléchissant       Airbag
##   Indemne                     0.745808231         0.019960804  0.012702330
##   Mort                        0.403852128         0.186393290  0.062131097
##   Blessé hospitalisé          0.314383023         0.183390097  0.047157453
##   Blessé léger                0.652077166         0.044503783  0.011609682
##
##                               Gants Gants + Airbag Non déterminable        Autre
##   Indemne              0.016331567    0.000000000      9.221891558  0.081657836
##   Mort                 0.341721031    0.000000000     12.364088226  0.186393290
##   Blessé hospitalisé   0.377259628    0.000000000     10.206968824  0.162431229
##   Blessé léger         0.042568836    0.005804841     12.234670382  0.052243571
##
##                      Non renseigné
##   Indemne              4.837773100
##   Mort                 0.186393290
##   Blessé hospitalisé   0.083835473
##   Blessé léger         0.090942513
```

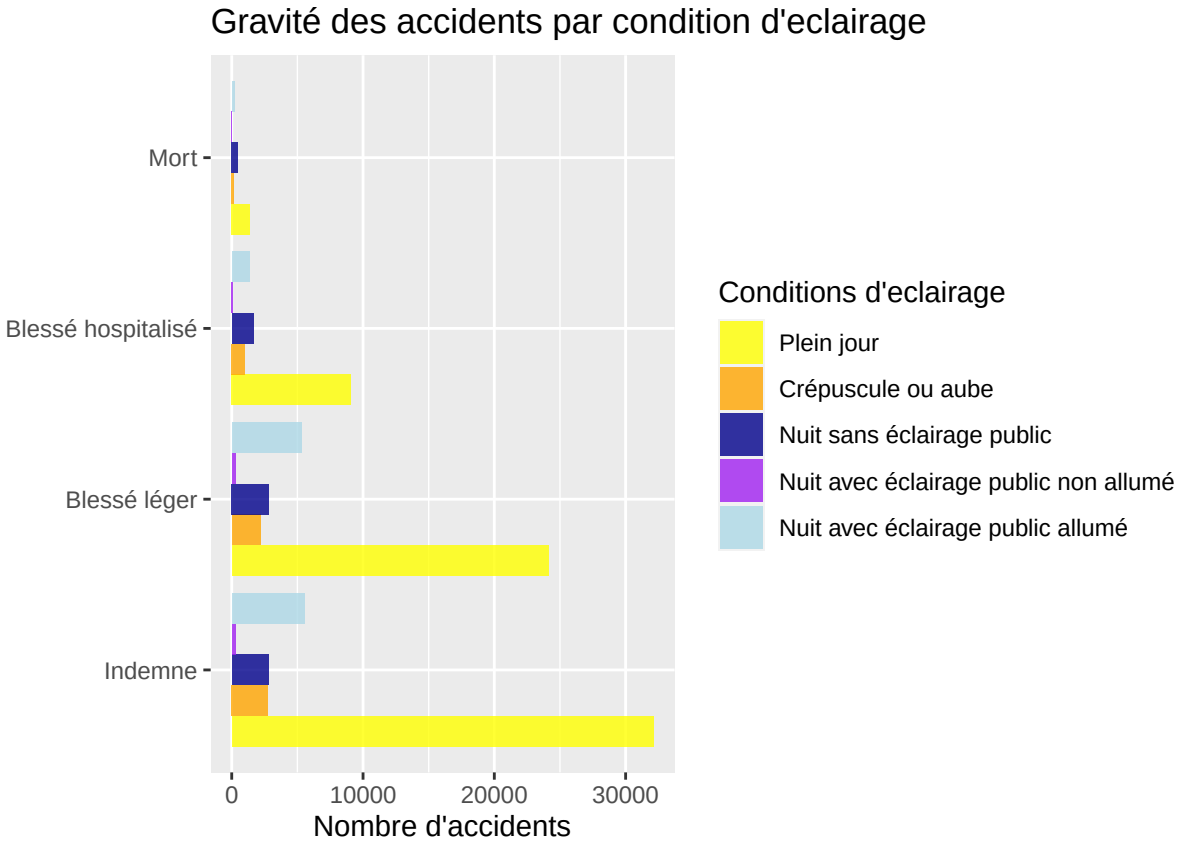
Dans les faits, 79.8% des accidents où les passagers ont porté une ceinture de sécurité, ils sont sortis indemnes.

On peut tirer plusieurs conclusions de ce tableau, par exemple :

- Dans l'ensemble des accidents, la majorité des passagers portent une ceinture de sécurité.
- Les passagers qui ne portent aucun équipement de sécurité ont un risque accru d'être gravement blessés ou tués dans un accident.
- Le port du casque est associé à une gravité plus élevée des accidents, probablement en raison du fait que les accidents où le casque est porté sont principalement des accidents de moto.
- Les passagers qui portent un gilet réfléchissant ou un dispositif pour enfants ont des taux de mortalité et de blessures graves relativement faibles.
- Les valeurs "Non déterminable" et "Non renseigné" pour l'équipement de sécurité sont relativement élevées, ce qui peut indiquer des lacunes dans la collecte de données ou des erreurs dans la déclaration des équipements de sécurité portés.

```
- On regarde le nb et la gravité en fonction de l'état de la route (mouillée, eneigée, sèche)
- On regarde le nb et la gravité en fonction de l'éclairage de la route (nuit noire, jour, éclairag
```

Est ce que la luminosité influence la gravité des accidents ?



On effectue un test du Chi² pour tester la corrélation entre la gravité et l'état de la route

```
##
##  Chi-squared test for given probabilities
##
## data:  df$grav
## X-squared = 97473, df = 129092, p-value = 1
```

La p-value élevée de 1 indique que l'hypothèse nulle, selon laquelle il n'y a pas de corrélation entre les deux variables, ne peut être rejetée. Autrement dit, il n'y a pas de preuve statistique pour affirmer qu'il y a une corrélation significative entre la gravité des blessures (variable "grav") et l'état de la surface (variable "surf") dans les accidents de la route en France en 2021.

Conclusion

Les conclusions que l'on peut tirer de cette étude sont les suivantes :

- La fin du confinement en juin 2021 a entraîné une augmentation du nombre d'accidents de la route en France, probablement en raison de l'augmentation du nombre de voitures sur les routes.
- Les accidents sont plus fréquents l'après-midi et le matin, pendant les périodes de pointe de la circulation, lorsque de nombreuses personnes vont au travail ou emmènent leurs enfants à l'école.

- L'âge des conducteurs est significativement associé à la gravité des accidents, avec une diminution de la probabilité d'avoir une gravité plus élevée pour chaque année supplémentaire. Les conducteurs indemnes ont tendance à être plus âgés que les autres groupes de gravité, tandis que les conducteurs blessés légèrement ont tendance à être plus jeunes.
- Le sexe ne semble pas influencer significativement la gravité des accidents, bien que le nombre total de conducteurs soit plus élevé chez les hommes que chez les femmes.

Ces conclusions peuvent être utilisées pour élaborer des politiques visant à améliorer la sécurité routière, telles que la sensibilisation aux dangers de la conduite pendant les périodes de pointe de la circulation, la formation des conducteurs plus jeunes et l'amélioration de la sécurité des routes pour les conducteurs plus âgés.