

**Rapport d’une analyse statistique en R**.

**BAC+3 (Filière : Ingénierie Logicielle)**

Sous le Thème

Réalisation d’une étude en langage R sur les bases de données annuelles des accidents corporels entre 2005-2018 et 2019 en France.

**Réalisés par :**

**Jihane Karroum**

**Introduction :**

Ce projet consiste à réaliser une étude statistique en langage R sur les bases de données annuelles des accidents corporels entre 2005-2018 et 2019 en France.

Pour chaque accident corporel (soit 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), des saisies d’informations décrivant l’accident sont effectuées par l’unité des forces de l’ordre (police, gendarmerie, etc.) qui est intervenue sur le lieu de l’accident. Ces saisies sont rassemblées dans une fiche intitulée bulletin d’analyse des accidents corporels. L’ensemble de ces fiches constitue le fichier national des accidents corporels de la circulation dit "Fichier BAAC" administré par l’Observatoire national interministériel de la sécurité routière "ONISR".

Les bases de données, extraites du fichier BAAC, répertorient l'intégralité des accidents corporels de la circulation intervenus durant une année précise en France métropolitaine ainsi que les départements d’Outre-mer (Guadeloupe, Guyane, Martinique, La Réunion et Mayotte depuis 2012) avec une description simplifiée. Cela comprend des informations de localisation de l’accident, telles que renseignées ainsi que des informations concernant les caractéristiques de l’accident et son lieu, les véhicules impliqués et leurs victimes.

L’enjeu de ce projet est de comparer les résultats obtenus entre 2005 jusqu’à 2018 et 2019 et en tirer des conclusions.

La réalisation de ce projet a passé par les étapes suivantes :

* Le téléchargement des bases de données depuis le site.
* L’importation des données en un fichier R.
* Le nettoyage des donnés et la vérification de persistance des données.
* La réalisation de différent test statistique sur les données.
* Représentation des résultats sous forme de graphes.
* Tirer des conclusions.

*Le téléchargement des bases de données :*

Nous avons importé les données depuis le site :

<https://www.onisr.securite-routiere.gouv.fr/en/data-tools/open-data>

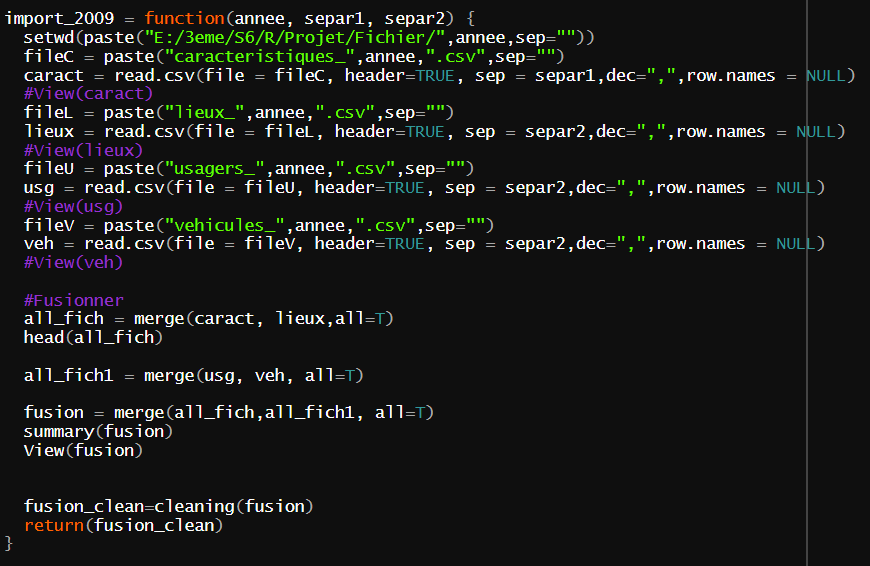
Les bases de données sont sous forme de quatre fichiers (caracteristique.csv, lieux.csv, usagers.csv et vehicules.csv) pour chaque année de 2005 jusqu’à 2019, plus un fichier PDF qui explique les bases de données.

*L’importation des données en R :*

Vu que nous disposons de plusieurs fichiers donc nous avons importé les données grâce a une fonction qui nous a permis de récupérer tous les fichiers en précisant l’année et le séparateur, par la suite elle fusionne les quatre fichiers d’une année dans un seul dataframe, enfin, elle fait appel à la fonction de nettoyage qui traite les valeurs manquantes.



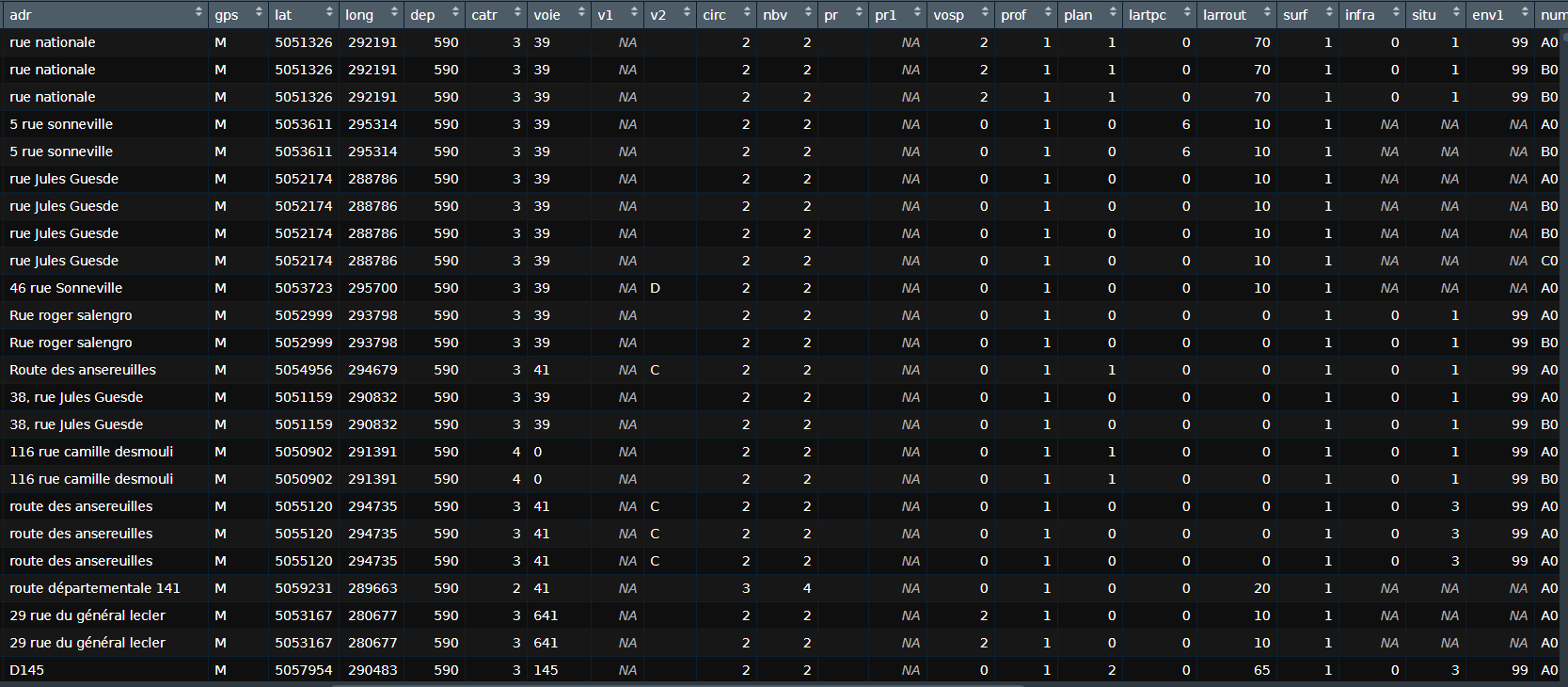
Pour l’année 2009, les séparateurs des fichiers en question ne sont pas identiques, donc nous avons créé une autre fonction qui traite ce cas exceptionnel.

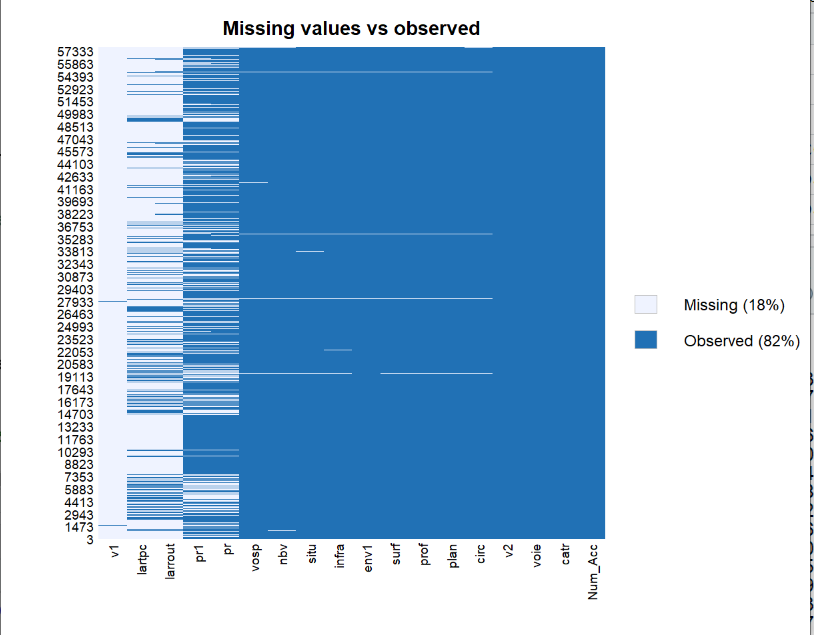


*Le nettoyage de données :*

Avant d’utiliser notre base de données, nous avons besoin de nettoyer les données en remplaçant les données manquantes.

Nous avons affiché le résultat de notre dataframe grâce à la méthode prédéfinis SUMMARY, en plus, nous avons utilisé le graphe suivant qui nous a permis de visualiser le pourcentage des données manquantes.





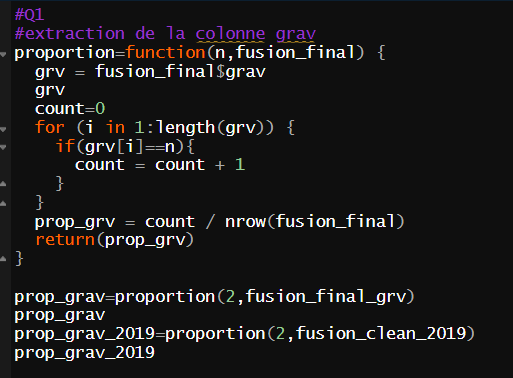
Nous nous sommes basés sur le résultat de ce graphe et nous avons créé une fonction qui traite les données manquantes en les remplaçant par zéro si leur % est inférieur à 10% sinon on les remplace par la moyenne.

*La réalisation des tests :*

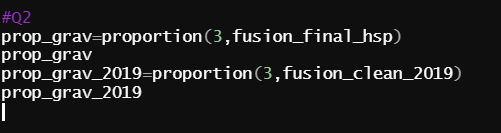
Les tests réalisés ont pour but de répondre aux questions suivantes :

1. Est-ce que c’est vrai que la proportion *d’accidents mortels* (impliquant au moins un décès) est augmentée en 2019 ?
2. Même question qu’au point précèdent mais par rapport aux nombre de *blesses hospitalisé*.
3. Si l’on veut intervenir par une manœuvre soit au niveau législatif soit au niveau de contraste local (contrôles policiers, etc.) on pourrait chercher à comprendre si la *catégorie de véhicule* est significative (vous considérerez seulement trois catégories : deux roues, voitures, poids lourds). Mettez en place un test qui pourrait répondre à cette question.
4. Toujours en allant dans le sens de la question précédente cherchez à comprendre si *l’autoroute* est plus dangereuse qu’une *route nationale* ou une *route départementale*.
5. Pour terminer, cherchez à comprendre sur la base des données à votre disposition si le vieux préjugé que les conducteurs *femmes* sont plus dangereuses que les *hommes* a un quelconque fondement ou si l’on devrait plutôt l’énoncer dans l’autre sens (les hommes sont plus dangereux que les femmes).

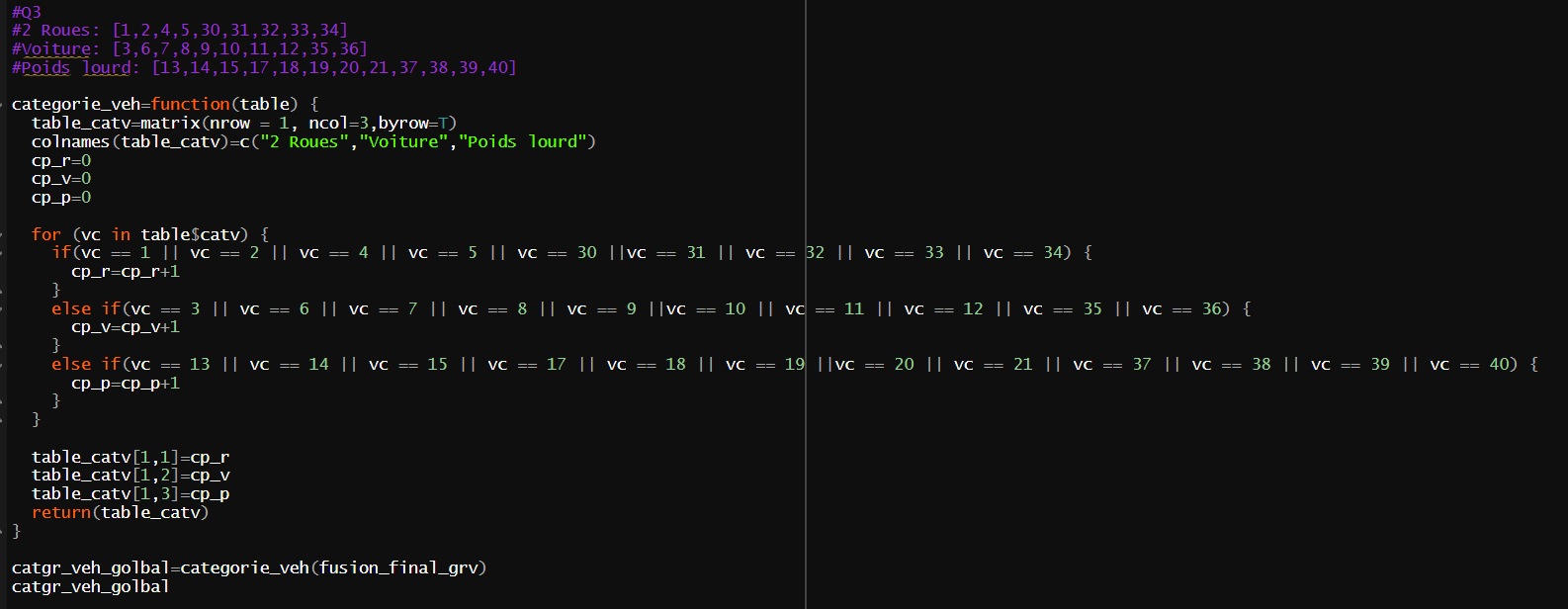
**Pour la première question :** nous avons créé une fonction qui sert à récupérer le nombre d’accidents mortels de 2005 jusqu’à 2018 et on l’a comparé avec celui de 2019.



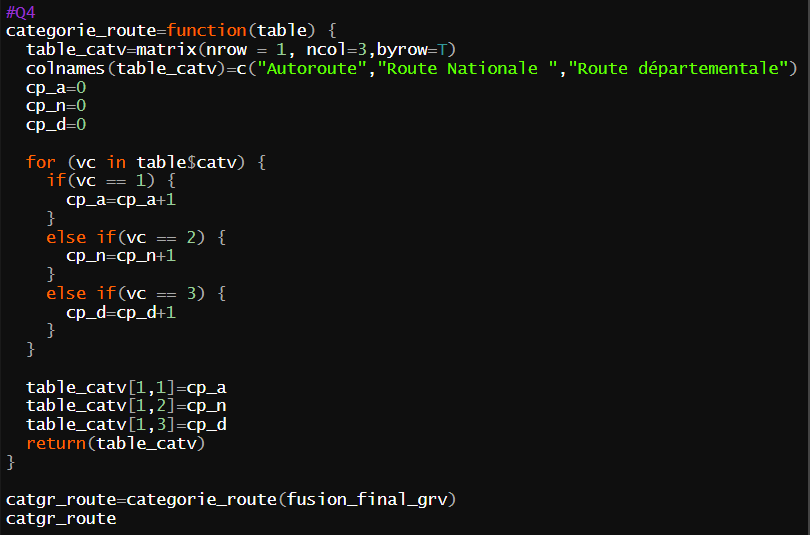
**Pour la deuxième question :** nous avons mis en place une fonction qui nous a permis de récupérer le nombre de blessés hospitalisés de 2005 jusqu’à 2017 et on l’a comparé avec celui de 2019, nous n’avons pas pris les données de 2018 pour respecter les consignes du fichier de description.



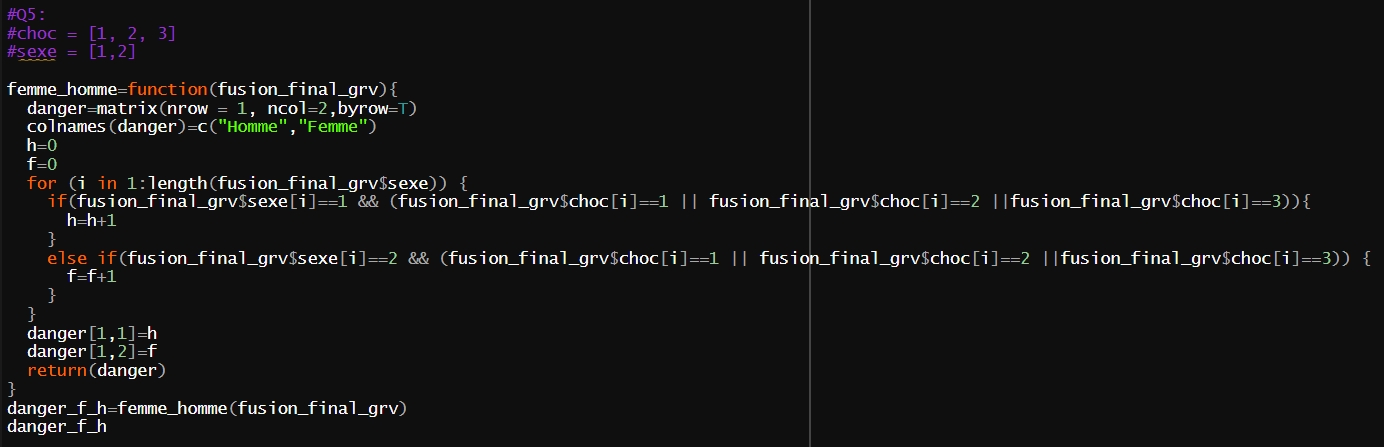
**Pour la troisième question :** nous avons effectué une fonction qui récupère les catégories de véhicules et teste la signification de ces derniers par rapport au nombre total d’accidents.



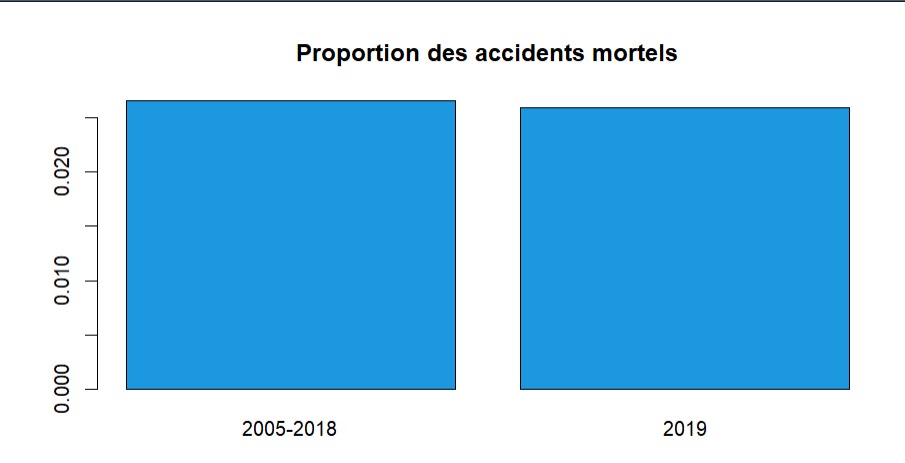
**Pour la quatrième question :** nous avons créé une fonction qui permet de savoir la route la plus dangereuse entre autoroute, route nationale ou route départementale.



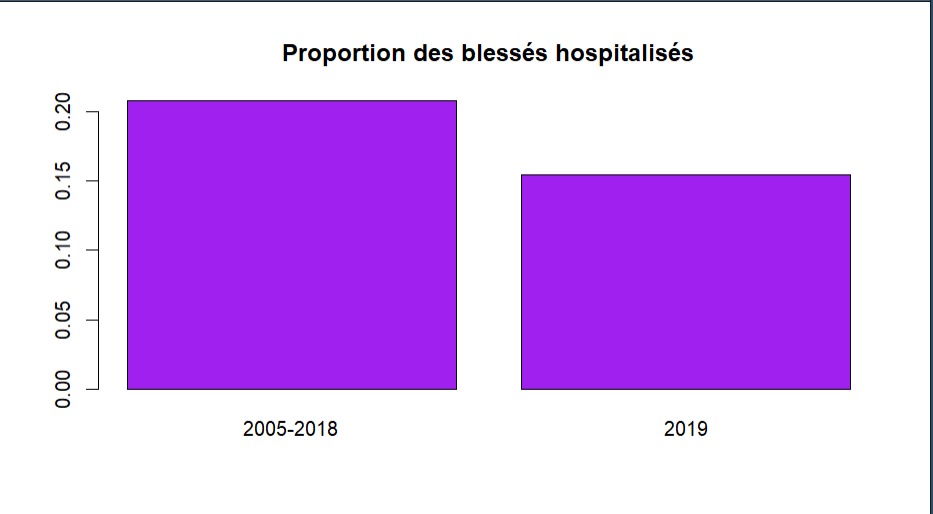
**Pour la cinquième question :** nous avons récupérer le nombre d’accident par sexe et en fonction de choc (avant, avant gauche, avant droit).



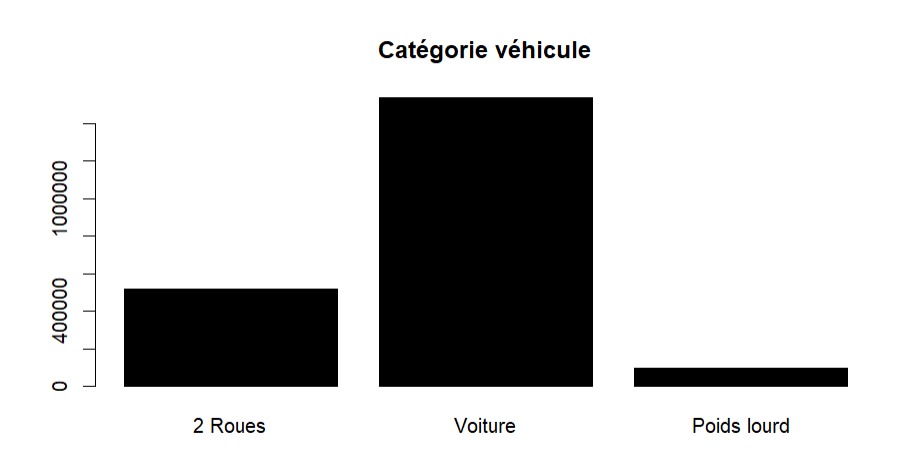
*Représentations des résultats sous forme de graphes :*

*1-*

Nous constatons qu’il y a une diminution de la proportion des accidents mortels entre 2005-2018 et 2019.

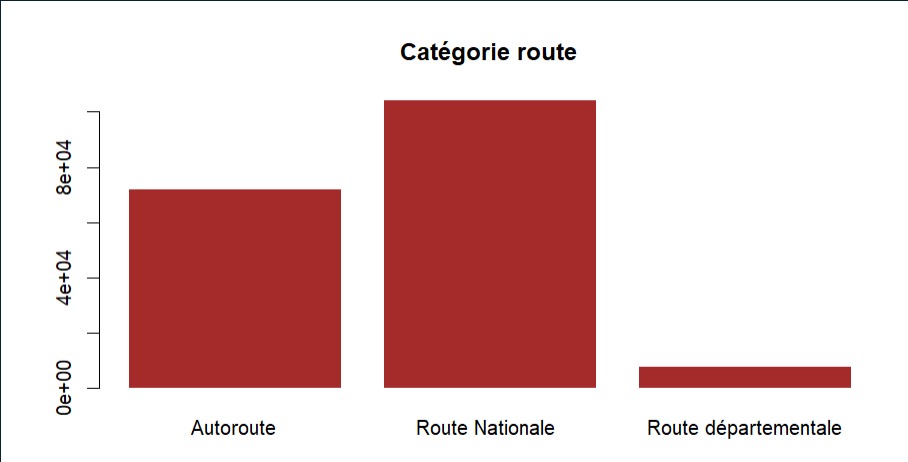
*2-*

Ce diagramme nous montre que le nombre des blessés hospitalisés à diminuer d’un pourcentage important.

*3- *

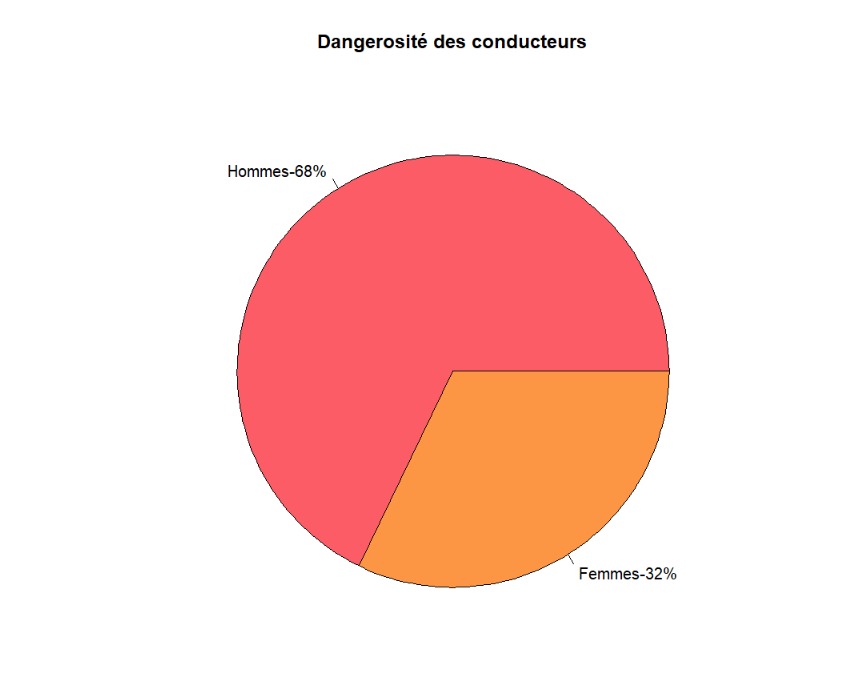
Le nombre d’accidents avec voitures est très élevés par rapport à celui des véhicules a deux roues et les véhicules poids lourds.

*4-*

**

Le nombre d’accidents situés en route nationale est très grand par rapport aux accidents situés en autoroute qui est un nom important aussi, par contre nous constatons que le nombre d’accident en route départementale est faible.

*5-*

**

*Ce graphe nous permet de conclure que les hommes sont plus dangereux que les femmes.*