Rapport Projet 2022

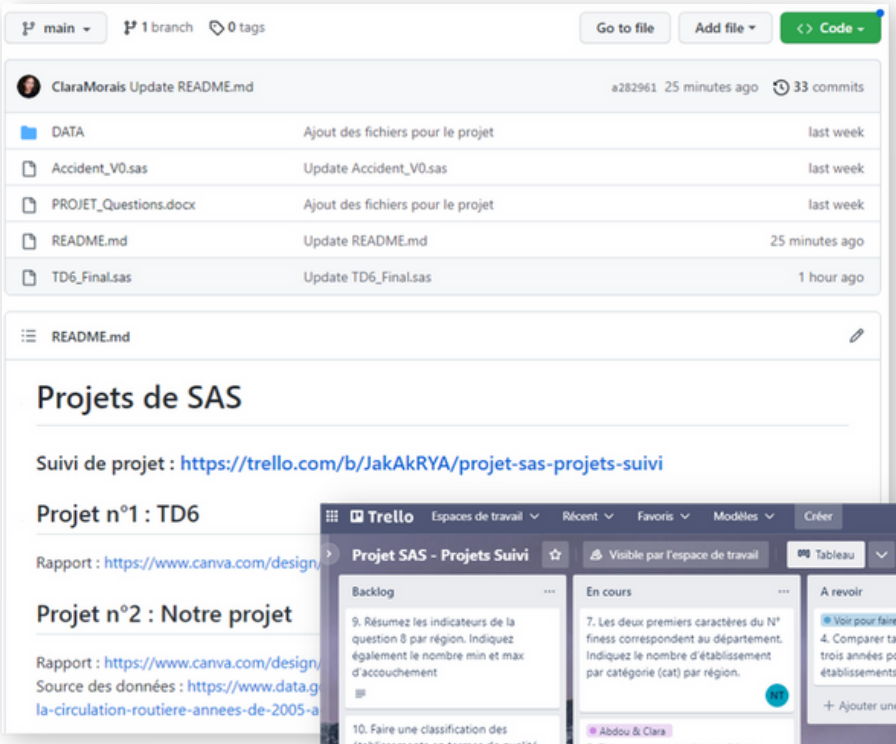
Abdourahmane NDIAYE
Quentin MONTALAND
Nathalie TANG
Clara MORAIS

Université Lumière Lyon 2
Institut de la communication

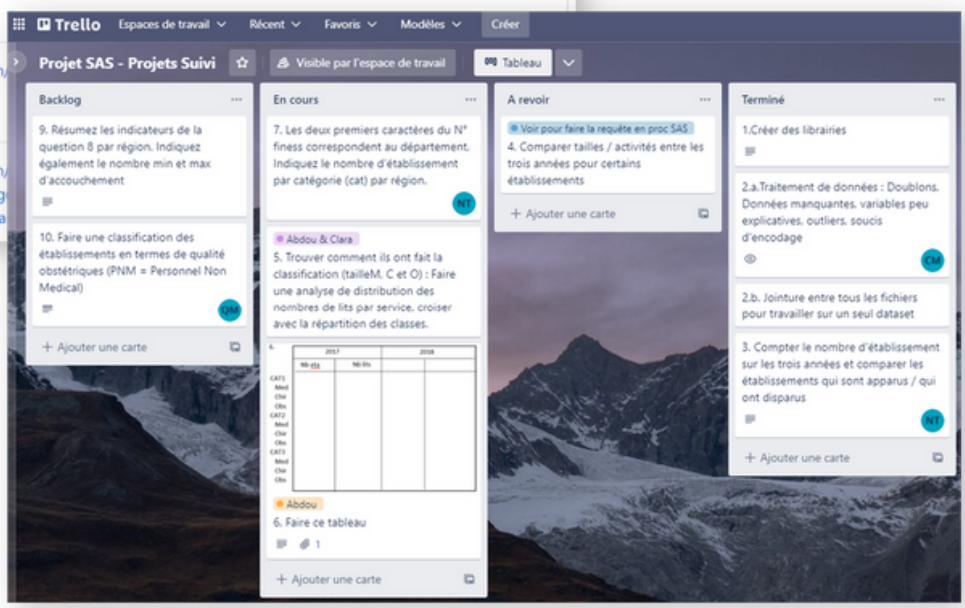
Sommaire

Organisation des projets

En ce qui concerne l'organisation des deux projets, nous avons suivi la même procédure. Tout d'abord, nous avons utilisé Discord pour communiquer et avons créé un groupe de discussion pour cela. Ensuite, afin de centraliser nos travaux (fichiers SAS, fichiers de données, ressources externes), nous avons utilisé GitHub, où chacun transmettait ses travaux pour que tout le monde puisse avoir la dernière version. Pour la répartition des tâches, chaque personne choisissait les questions auxquelles elle souhaitait répondre et l'avancement de celles-ci était organisé sur un Trello. Par ailleurs, le choix de former un groupe de 4 se justifie par rapport au nombre important de projet que nous avons à rendre en parallèle de ceux de l'enseignement de SAS.



Aperçu de notre page GitHub



Aperçu de notre page Trello

Introduction

Présentation du jeu de données

Pour notre projet, nous avons décidé de travailler sur la base de données annuelle des accidents corporels de la circulation routière pour 2021. Elle est accessible via le site data.gouv.fr et représentent l'intégralité des données relatives aux accidents (mortels ou non) de la route relevé par les forces de l'ordre. Ces dernières sont réparties en quatre fichiers :

- La rubrique **caractéristiques** qui décrit les circonstances générales de l'accident
- La rubrique **lieux** qui décrit le lieu de l'accident
- La rubrique **véhicules**, présentant certaines caracteristiques du véhicule impliqué lors de l'accident mais aussi des éléments d'environnement
- La rubrique **usagers** qui recense les informations relatives aux personnes impliquées dans l'accident

Source des fichiers : *https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2021/#description*

L'idée générale de ce projet est de mener une analyse sur ces fichiers tout en essayant de parcourir les possibilités principales que nous offre SAS. Tout d'abord, nous verrons comment importer et nettoyer des données. Puis, des premières analyses descriptives seront faites avant d'entamer des analyses statistiques plus précises. Enfin, vous pourrez trouver en Partie 4 du programme SAS des analyses supplémentaires qui ne seront pas détaillées dans ce rapport.

Questions

PARTIE 1 : Préparation des données

1. Importer les fichiers et les traiter pour n'obtenir qu'un jeu de données

2. Nettoyer les données : remplacement des valeurs nulles, changement des types, repérer les doublons et les valeurs aberrantes

PARTIE 2 : Analyses descriptives

1. Afficher un tableau montrant le nombre d'accident par département et par localisation (Hors agglomération ou en agglomération). Quelles sont les analyses que l'on peut tirer de ce tableau ?

2. En se basant sur des résultats graphiques, est-ce que les conditions d'environnement ont un impact sur le nombre d'accidents ? (éclairage, météo, état de la route).

3. Représenter graphiquement les lieux où se sont produits les accidents.

PARTIE 3 : Analyses inférentielles

1. Réaliser un test d'indépendance du Khi-Deux afin de déterminer si la gravité d'un accident est plus importante pour une catégorie d'usager. Quelles conclusions peut-on tirer de ce test ?

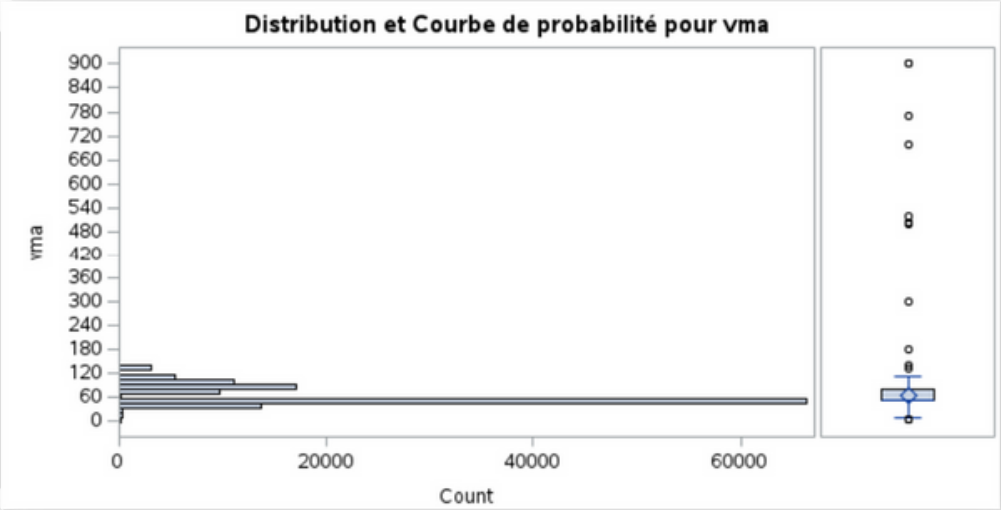
2. Réaliser un test d'ANOVA avec la variable de votre choix. Quelles conclusions peut-on tirer de ce test ?

Partie 1 - Question 1

Cf. Programme SAS

Partie 1 - Question 2

Pour améliorer la cohérence des données et obtenir de meilleures performances, nous avons converti certaines variables en données numériques (cf. bibliographie). Ensuite, après avoir éliminé les éventuels doublons, nous avons analysé ces variables numériques pour détecter des outliers. Tout d'abord, nous avons remarqué que dans la variable **vma** (vitesse maximale autorisée lors de l'accident), un certain nombre de valeurs aberrantes qui semblent être des erreurs de saisie font leur apparition (des valeurs supérieures à la limite maximale autorisée en France, soit 130 km/h).



En outre, la variable **lartpc** (largeur du terre-plein central en m) présente aussi une valeur extrême de 40m. Or, après avoir effectué une analyse plus approfondie, on remarque que l'accident à eu lieu sur un giratoire. On peut donc supposer que les 40m correspondent à la largeur de celui-ci.

Observations extrêmes			
La plus petite		La plus grande	
Valeur	Obs	Valeur	Obs
0	128797	12	121056
0	128408	13	126324
0	128407	13	126325
0	128406	40	116518
0	128405	40	116519

	Nb_accident		Total
	Agglomeration		
	1	2	
01	612	389	1001
02	223	200	423
03	322	195	517
04	269	155	424
05	249	318	567
06	563	1553	2116
07	322	229	551
08	94	105	199
09	199	183	382
10	259	603	862

20 201 Hors agglomération (1) ACCIDENTS

36 317 En agglomération (2) ACCIDENTS

Départements avec le plus d'accidents

dep	nb_accidents
75	5069
93	2896
13	2779
94	2493
69	2443

Partie 2 - Question 2

% d'accidents en fonction de la météo

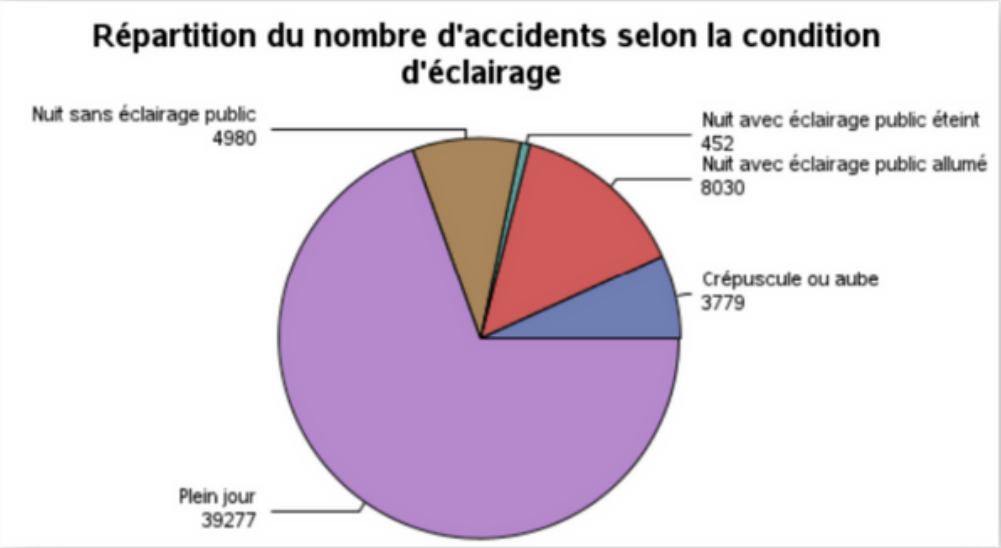
Conditions météorologiques									
.	1	2	3	4	5	6	7	8	9
0.01	79.76	10.62	2.05	0.49	0.69	0.20	1.76	3.92	0.49

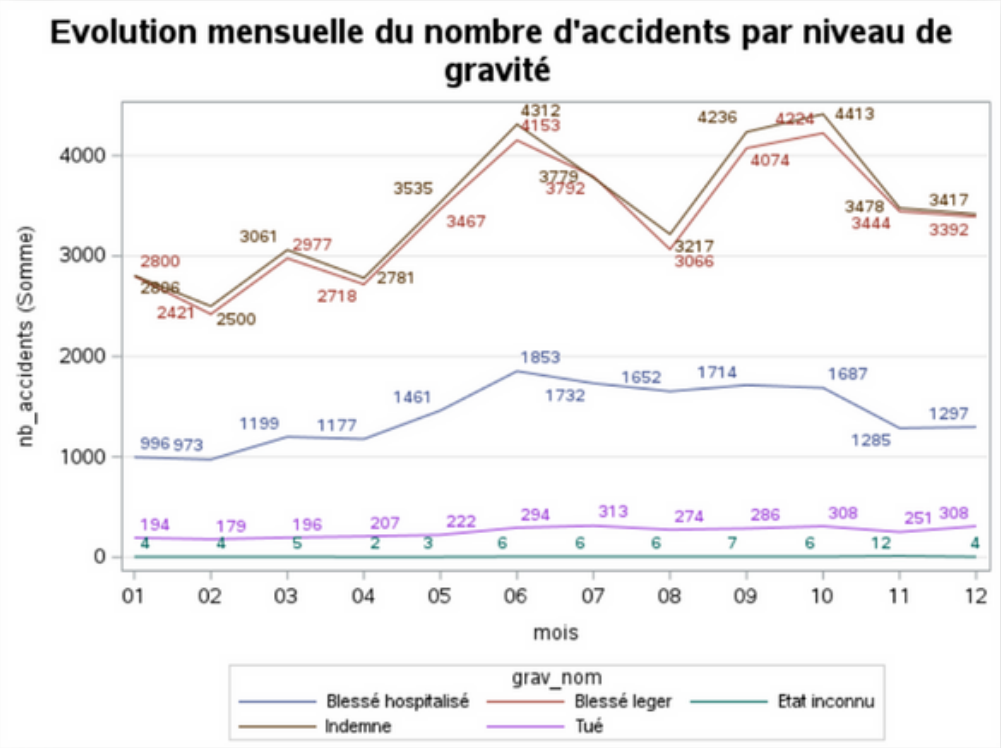
1 - Normale **1ère position**
2 - Pluie légère **2e position**
3 - Pluie forte
4 - Neige - grêle
5 - Brouillard / fumée
6 - Vent fort / tempête
7 - Temps éblouissant
8 - Temps couvert **3e position**
9 - Autre

% d'accidents en fonction de l'état de la route

Etat de la route									
.	1	2	3	4	5	6	7	8	9
0.02	80.82	17.69	0.12	0.04	0.27	0.06	0.46	0.10	0.40

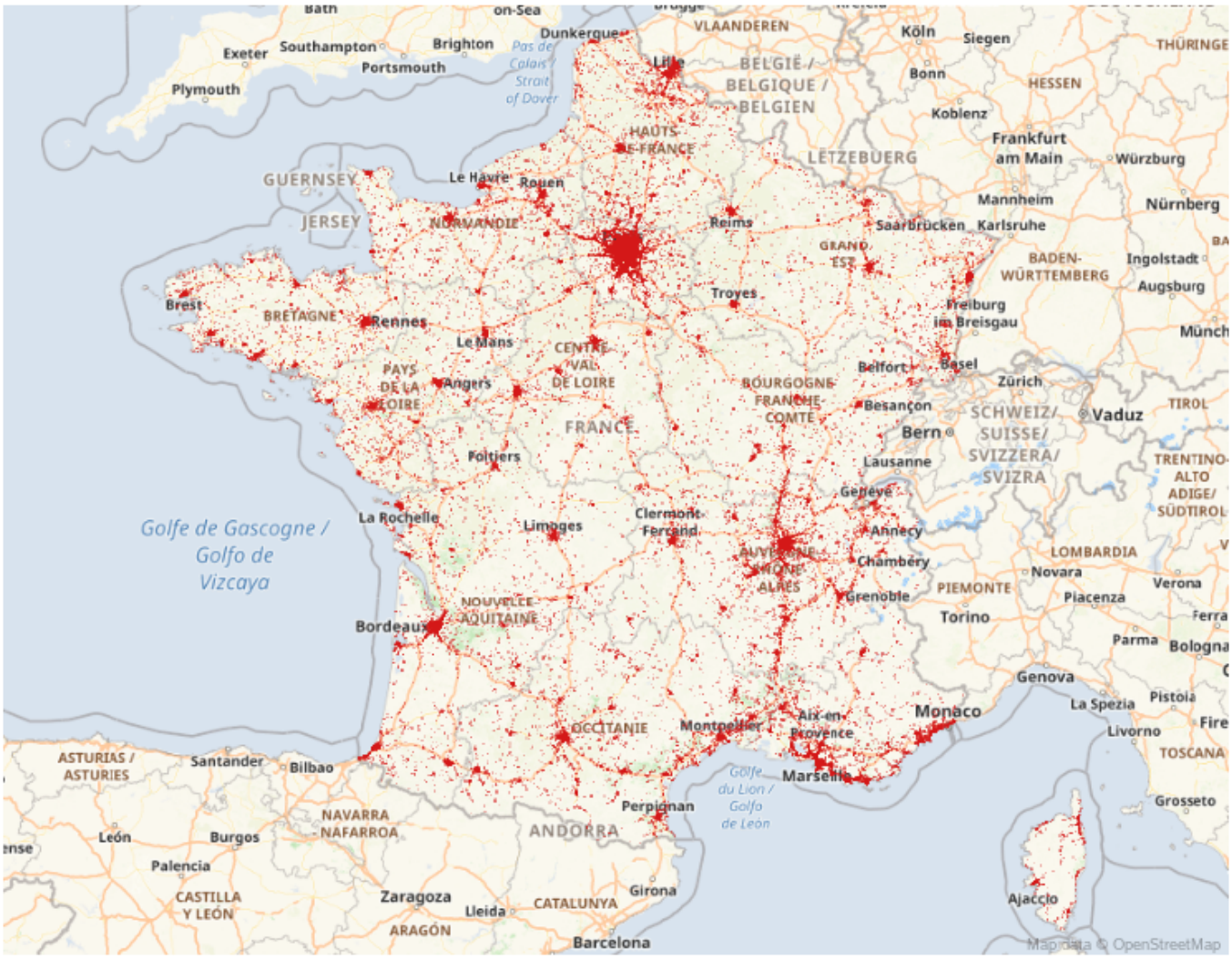
1 - Normale **1ère position**
2 - Mouillée **2e position**
3 - Flaques
4 - Inondée
5 - Enneigée
6 - Boue
7 - Verglacée
8 - Corps gras / huile
9 - Autre





Grâce à nos diverses visualisations, nous pouvons constater plusieurs choses en ce qui concerne les conditions des accidents. Tout d'abord, il est difficile de dire si les conditions météorologiques ainsi que l'état de la route ont un impact important. En effet, la majorité des accidents ont lieu par temps "neutre" sur une route en bon état. De même, nos données ne permettent pas de conclure qu'il y a davantage d'accidents sur des routes mal éclairées, comme le montre le graphique en secteur.

Cartographie des accidents de voiture



TOP 3 DES COMMUNES AVEC LE PLUS D'ACCIDENTS
En région parisienne

1er - Paris 16e

2e – Paris 12e

3e – Paris 19e

TOP 3 DES COMMUNES AVEC LE PLUS D'ACCIDENTS
Hors région parisienne

1er - Cayenne

2e – Rennes

3e – Toulouse

Nous allons maintenant nous intéresser aux tests statistiques que nous propose SAS. Par exemple, nous avons souhaité savoir si le fait d'être passager ou conducteur est lié à la gravité de l'accident. En d'autres termes, est-ce que les deux variables sont-elles indépendantes ou non. Pour cela, nous avons fait appel au test du Khi-deux qui permet de tester l'indépendance entre deux variables qualitatives. Nous avons en premier lieu créé un nouveau dataset où nous avons sélectionner les modalités qui nous intéressaient dans les variables. On obtient les analyses suivantes :

H0 : Il existe un lien entre les deux variables
H1 : Il n'existe pas de lien

Statistiques pour la table de catu_nom par cate_grav

Statistique	DDL	Valeur	Prob
Khi-2	1	0.2540	0.6143
Test du rapport de vraisemblance	1	0.2536	0.6146
Khi-2 continuité ajustée	1	0.2439	0.6214
Khi-2 de Mantel-Haenszel	1	0.2540	0.6143
Coefficient Phi		-0.0015	
Coefficient de contingence		0.0015	
V de Cramer		-0.0015	

Fréquence
Attendu
Ecart
Pourcentage

Table de catu_nom par cate_grav

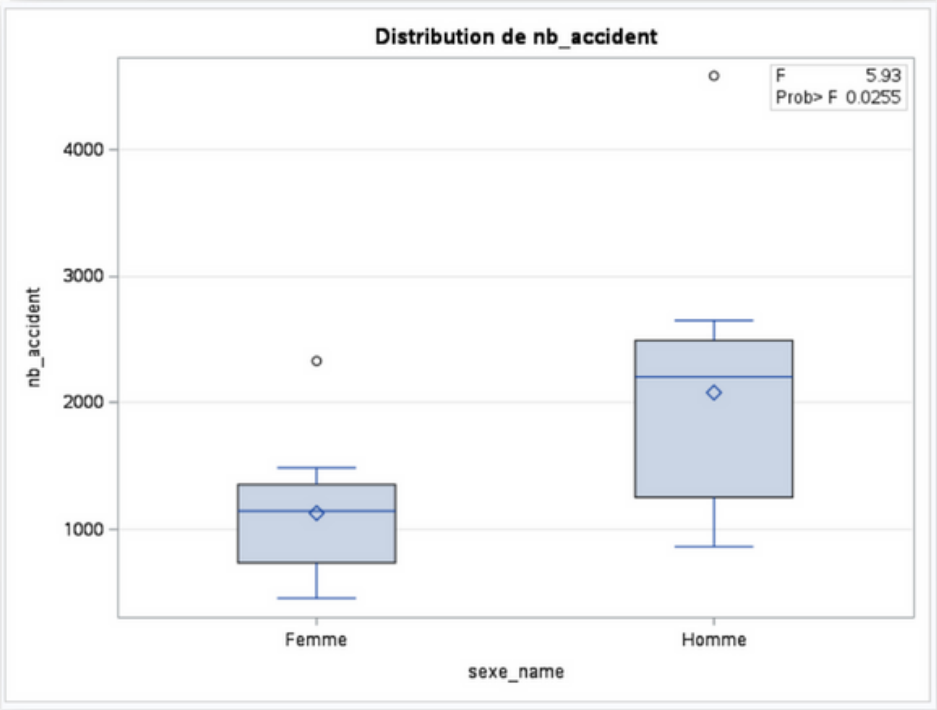
catu_nom	cate_grav		
	Accident grave	Accident leger	Total
Conducteur	15623 15648 -24.99 13.06	81593 81568 24.99 68.21	97216 81.27
Passager	3631 3606 24.99 3.04	18772 18797 -24.99 15.69	22403 18.73
Total	19254 16.10	100365 83.90	119619 100.00

A première vue, la valeur de notre p-value (encadré violet) laisse entendre qu'il n'existe pas de lien significatif entre nos deux variables puisqu'elle est supérieure à notre seuil (0,05). Toutefois, grâce à l'option "EXPECTED" de la procédure Freq, on remarque que nos effectifs théoriques ("attendu" dans la tableau) sont nuls. Cette information nous indique que la condition de répartition uniforme de nos données n'est pas respectée et par conséquent, on ne peut pas affirmer que nos deux variables sont indépendantes.

A présent, nous allons utiliser la PROC ANOVA de SAS pour réaliser un test de l'ANOVA, afin de savoir s'il existe une différence significative entre le nombre d'accidents causés par des hommes ou par des femmes. Pour ce faire, comme pour le test du Khi-Deux, nous avons créé un dataset prévu à cet effet, en sélectionnant les dix départements où il y a le plus d'accidents. Une fois la PROC ANOVA lancée, on obtient les résultats suivants :

H0 : Le nombre d'accident ne diffère pas selon le sexe
H1 : Il existe une différence

Source	DDL	Anova SS	Carré moyen	Valeur F	Pr > F
sexe_name	1	4501107.200	4501107.200	5.93	0.0255



Ainsi, sur la base de notre échantillon, étant donné que notre p-value est inférieure à notre seuil (0,05), on rejette l'hypothèse H0 et on conclu qu'il existe une différence significative entre les accidents causés par des hommes ou par des femmes.

Bibliographie

1. *Convert multiple variables between character and numeric formats in SAS.* (2022, 9 février). The Chemical Statistician. https://chemicalstatistician.wordpress.com/2018/04/27/convert-multiple-variables-between-character-and-numeric-formats-in-sas/comment-page-1/