PROJET BIG DATA

Tristan SAEZ - Vincent LE BRENN
Adrien LEBOUCHER
mai 2023

TABLE DES MATIERES

GESTION DE PROJET

01

LE PROJET

Données d'accidents corporels de la circulation routière

REPRÉSENTATIONS GRAPHIQUES & STATISTIQUES À PARTIR D'UN FICHIER CSV

DIAGRAMME DE GANTT

PROJECT BIG DATA

ISEN Yncrea Brest

Tristan SAEZ - Vincent
LE BRENN - Adrien
LEBOUCHER

SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

Project Start: Wed, 5/24/2023

Display Week: 1

May 22, 2023

TASK	ASSIGNED TO	PROGRESS	START	END	22 M	23 T	24 W	25 T	26 F	27 S	28 S
Préparation des données											
Gestion info manquantes	Vincent	100%	5/24/23	5/24/23							
Recoder var multimodale	Tristan	100%	5/24/23	5/24/23							
Formater var	Adrien	100%	5/24/23	5/24/23							
Serie chrono mensu et he	Tristan	100%	5/24/23	5/24/23							
Jeu de données accident	Vincent	100%	5/24/23	5/24/23							
Visualisation des données											
Représentation graph	Adrien	95%	5/25/23	5/25/23							
Histogrammes	Vincent	100%	5/25/23	5/25/23							
Carte géo accident	Tristan	100%	5/25/23	5/25/23							
"" accident grave	Tristan	100%	5/25/23	5/25/23							
Analyse des données											
Etude rela var qualitativ	Adrien	75%	5/26/23	5/26/23							
RegLineaire	Vincent	95%	5/26/23	5/27/23							
Bonus		0%	5/26/23	5/27/23							
Insert new rows ABOVE this one											

NETTOYAGE DES DONNÉES

02

Les données inexploitable

Suppression des données inexploitable
Exemple: Latitude 2009

	Num_Acc	num_veh	id_usa	date	ville	id_code_insee	latitude	l
41888	2.009E+11	D01	756283	5/29/2009 14:15	PARIS 19	75119	2009	
41889	2.009E+11	B01	756321	5/22/2009 15:30	PARIS 12	75112	2009	
41890	2.009E+11	C01	756322	5/22/2009 15:30	PARIS 12	75112	2009	
41891	2.009E+11	A01	756429	5/3/2009 14:00	PARIS 16	75116	2009	
41892	2.009E+11	B01	756430	5/3/2009 14:00	PARIS 16	75116	2009	
41893	2.009E+11	C01	756431	5/3/2009 14:00	PARIS 16	75116	2009	
41894	2.009E+11	A01	756545	5/1/2009 13:15	PARIS 16	75116	2009	
41895	2.009E+11	C01	756544	5/1/2009 13:15	PARIS 16	75116	2009	
41896	2.009E+11	A01	756551	5/1/2009 15:10	PARIS 19	75119	2009	
41897	2.009E+11	B01	756554	5/1/2009 15:10	PARIS 19	75119	2009	
41898	2.009E+11	C01	756552	5/1/2009 15:10	PARIS 19	75119	2009	
41899	2.009E+11	A01	756580	5/4/2009 13:15	PARIS 09	75109	2009	
41900	2.009E+11	A01	756682	5/7/2009 15:35	PARIS 19	75119	2009	
41901	2.009E+11	C01	756681	5/7/2009 15:35	PARIS 19	75119	2009	
41902	2.009E+11	A01	756695	5/7/2009 16:50	PARIS 14	75114	2009	
41903	2.009E+11	C01	756692	5/7/2009 16:50	PARIS 14	75114	2009	
41904	2.009E+11	D01	756693	5/7/2009 16:50	PARIS 14	75114	2009	
41905	2.009E+11	A01	756779	5/10/2009 12:30	PARIS 12	75112	2009	
41906	2.009E+11	B01	756778	5/10/2009 12:30	PARIS 12	75112	2009	
41907	2.009E+11	B01	756809	5/11/2009 8:30	PARIS 17	75117	2009	
41908	2.009E+11	C01	756811	5/11/2009 8:30	PARIS 17	75117	2009	
41909	2.009E+11	A01	756815	5/11/2009 8:55	PARIS 12	75112	2009	
41910	2.009E+11	B01	756814	5/11/2009 8:55	PARIS 12	75112	2009	

	descr_lum	descr_etat_surf	description_intersection	an_nais	age	place	des
60329	Nuit avec éclairage public allumé	Normale	Intersection en X	1987	36	1	Pré
60330	Plein jour	Normale	Hors intersection	1976	47	1	Util
60331	Plein jour	Normale	Hors intersection	1984	39	1	Pré
60332	Plein jour	Normale	Hors intersection	1956	67		Aut
60333	Nuit avec éclairage public allumé	Normale	Hors intersection	1969	54		Aut
60334	Plein jour	Normale	Hors intersection	1946	77	1	Util
60335	Plein jour	Normale	Autre intersection	1995	28	1	Pré
60336	Plein jour	Mouillée	Hors intersection	1968	55	1	Util
60337	Plein jour	Normale	Hors intersection	2007	16		Aut
60338	Plein jour	Normale	Hors intersection	1934	89		Aut
60339	Plein jour	Normale	Hors intersection	1973	50		Aut

Données 73 644 à 55 428

14 639 données avec une latitude de 2009
3967 données avec des colonnes vides

Difficulté: Connaître les fonctions qui simplifie le traitement de donnée
Syntaxe de R

Recodage des variables multimodales

OBJECTIF & MÉTHODE

- Permettre une analyse et une exploitation des données plus simple
- Gagner en performances et en vitesse d'analyse

- Récupérer chaque valeur connu dans la base
- Convertir les valeurs récupérées en valeurs numériques (1 à x, x étant le nombre de valeurs différentes)
- Modifier le jeu de données en remplaçant les variables multimodales par leur valeur numérique
- 4 valeurs pour la description de la gravité (1 à 4), 56 valeurs pour la description du véhicule (1 à 56)

Difficultés rencontrées

- Problème d'encodage de la base
- Problème d'indice des tableaux

Création des séries chronologiques

OBJECTIF & MÉTHODE

- Permettre une analyse des données plus poussée en fonction de paramètres pertinents
- Réaliser des statistiques sur les fréquences d'accidents et pouvoir, à terme, les prédire

- Récupérer la dates et le mois de chaque accident
- Créer 2 data.frame contenant les mois/semaines et une colonne pour les accidents
- Compter les accidents pour chaque semaine/ chaque mois et afficher les résultats

Difficultés rencontrées

- Récupération de la partie exploitable de la date (mois, date sans l'heure...)
- Calcul de la semaine de l'accident

	A	B	C	D	E
1	Column1	Code_Region	Region	descr_grav	accidents_100k
2	39	3	['GUYANE']	Indemne	0.207833127
3	38	3	['GUYANE']	Blessé hospitalisé	0.142885275
4	3	1	['GUADELOUPE']	Indemne	0.105705249
5	11	2	['MARTINIQUE']	Indemne	0.097120512
6	9	2	['MARTINIQUE']	Blessé hospitalisé	0.083568813
7	1	1	['GUADELOUPE']	Blessé hospitalisé	0.073032717
8	40	3	['GUYANE']	Tué	0.064947852
9	10	2	['MARTINIQUE']	Blessé léger	0.057594722
10	47	4	['REUNION']	Indemne	0.054320831
11	81	74	['LIMOUSIN']	Blessé hospitalisé	0.047028274
12	84	74	['LIMOUSIN']	Indemne	0.047028274
13	2	1	['GUADELOUPE']	Blessé léger	0.046125927
14	46	4	['REUNION']	Blessé léger	0.040982264
15	57	43	['FRANCHE-COMTE']	Blessé hospitalisé	0.032520465
16	59	43	['FRANCHE-COMTE']	Indemne	0.031985589
17	91	83	['AUVERGNE']	Indemne	0.031841362
18	15	21	['CHAMPAGNE-ARDENNE']	Indemne	0.029816913
19	82	74	['LIMOUSIN']	Blessé léger	0.029000769
20	31	25	['BASSE-NORMANDIE']	Indemne	0.026768316
21	37	3	['GUYANE']	Blessé léger	0.025979141
22	35	26	['BOURGOGNE']	Indemne	0.024162594
23	55	42	['ALSACE']	Indemne	0.023713056
24	21	23	['HAUTE-NORMANDIE']	Indemne	0.023474733
25	71	54	['POITOU-CHARENTES']	Indemne	0.022572798
26	90	83	['AUVERGNE']	Blessé léger	0.022452968
27	45	4	['REUNION']	Blessé hospitalisé	0.021651007
28	13	21	['CHAMPAGNE-ARDENNE']	Blessé hospitalisé	0.021619385

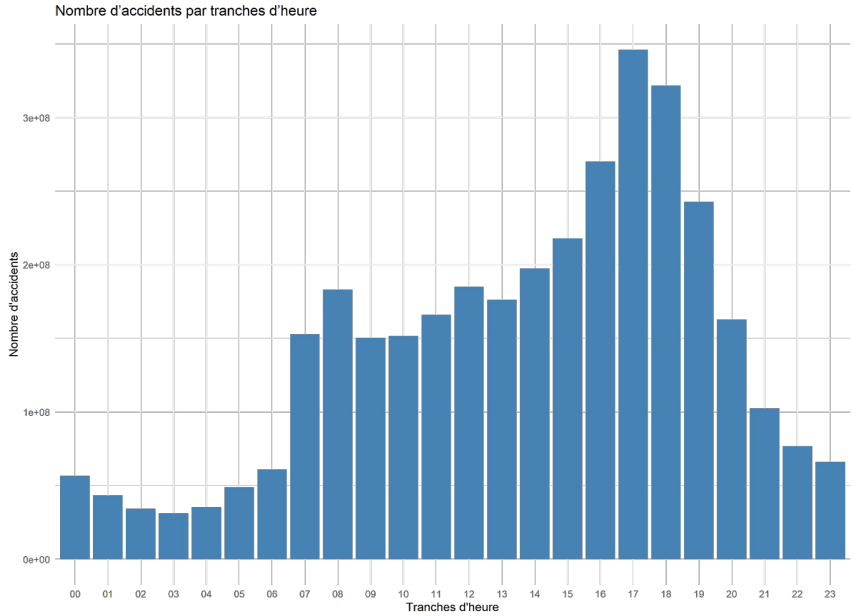
Merge :
en fonction du code insee
code de la región

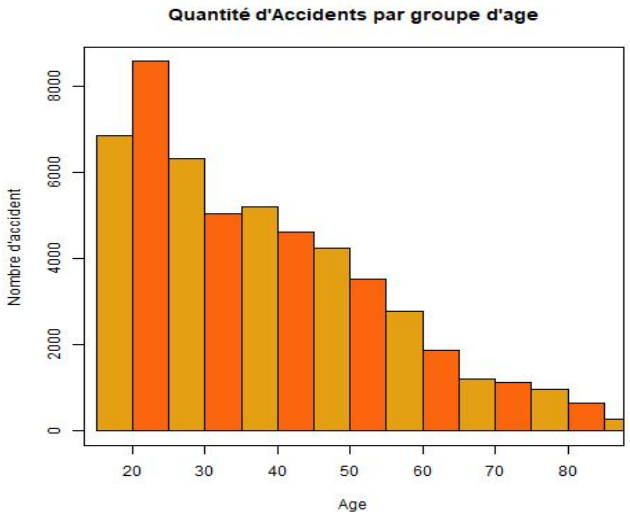
Agrege
en fonction de la région avec
lenght
puis par sum

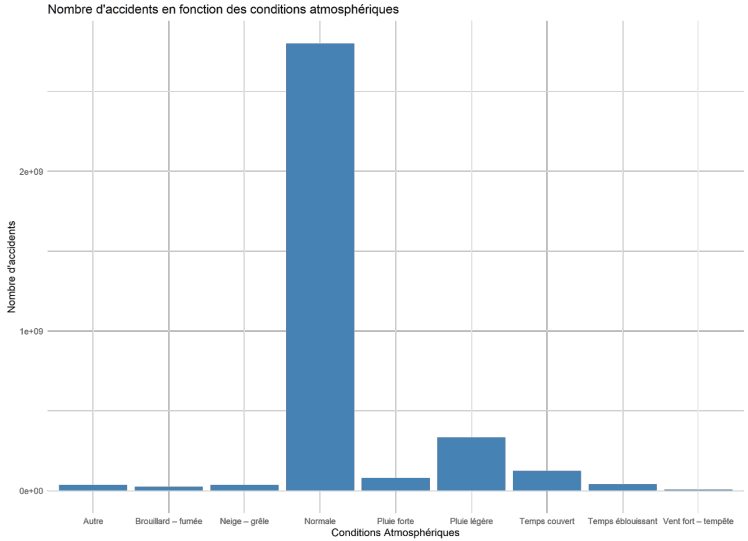
VISUALISATION DES DONNÉES

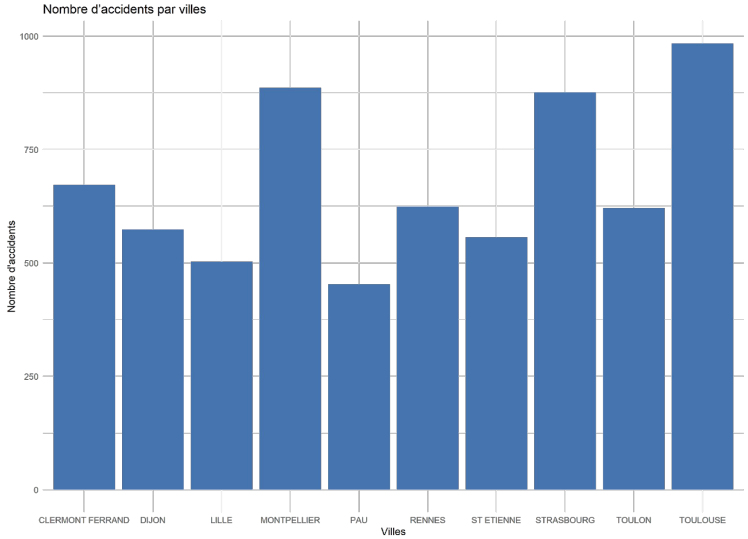
03

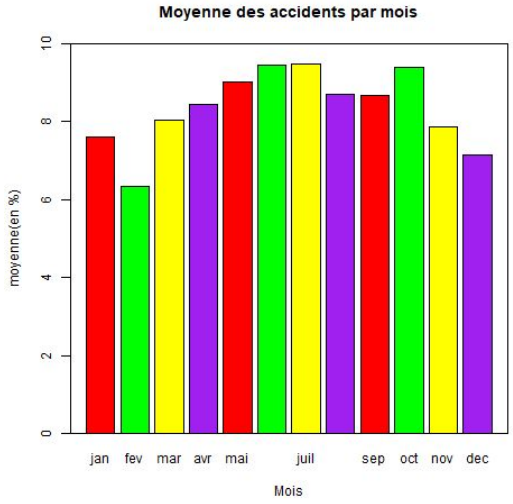
Des Histogrammes



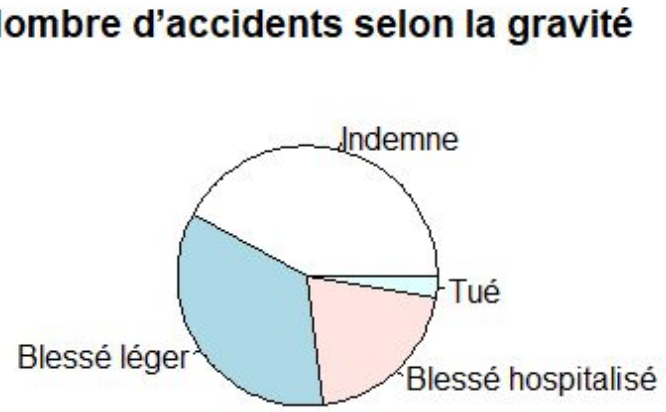

Quantité d'Accidents par groupe d'age
Nombre d'accident
0
2000
4000
6000
8000
20
30
40
50
60
70
80
Age


Nombre d'accidents en fonction des conditions atmosphériques
Nombre d'accidents
0e+00
1e+09
2e+09
Autre
Brouillard – fumée
Neige – grêle
Normale
Pluie forte
Pluie légère
Temps couvert
Temps éblouissant
Vent fort – tempête
Conditions Atmosphériques


Nombre d'accidents par villes
Nombre d'accidents
0
250
500
750
1000
CLERMONT FERRAND
DIJON
LILLE
MONTPELLIER
PAU
RENNES
ST ETIENNE
STRASBOURG
TOULON
TOULOUSE
Villes

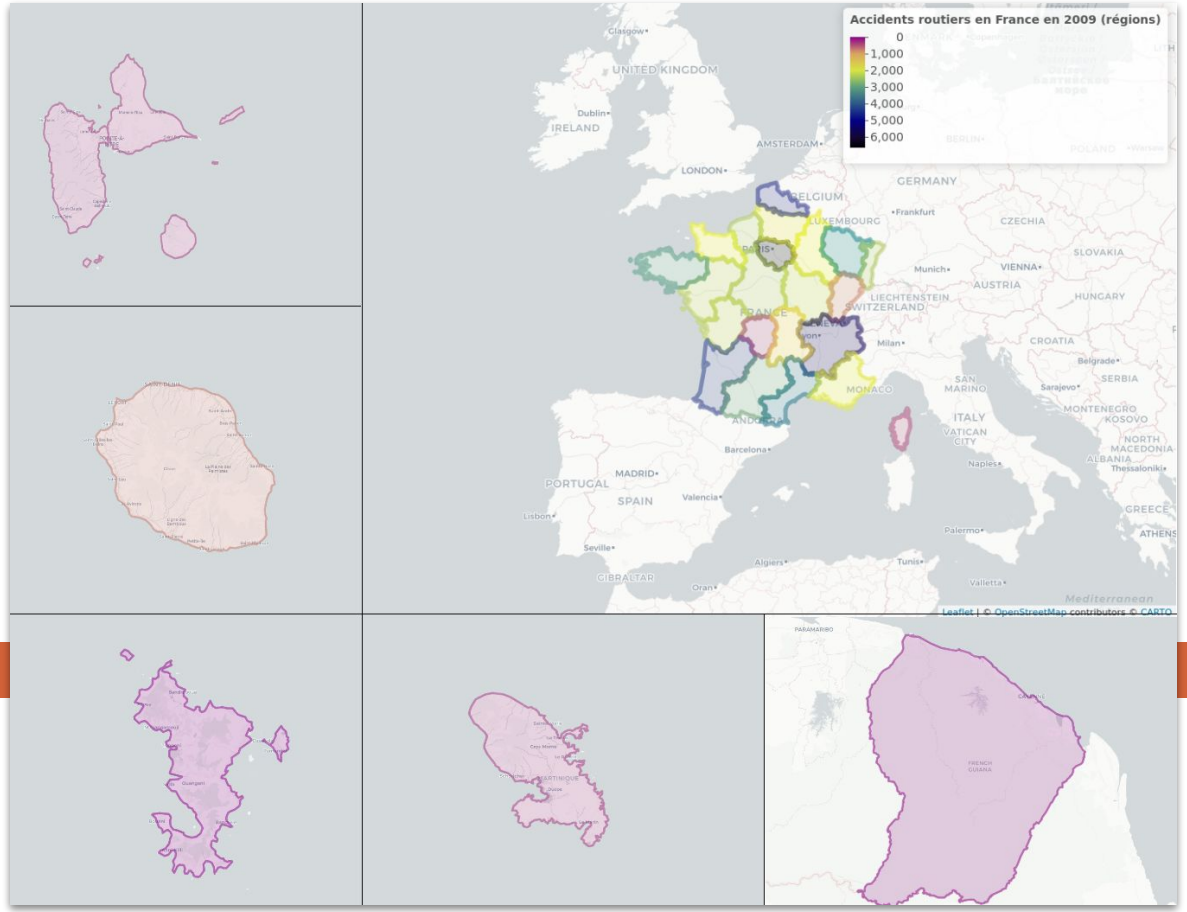

Moyenne des accidents par mois
moyenne(en %)
0
2
4
6
8
10
jan
fev
mar
avr
mai
juil
sep
oct
nov
dec
Mois

Des Diagrammes circulaires



Des Cartes

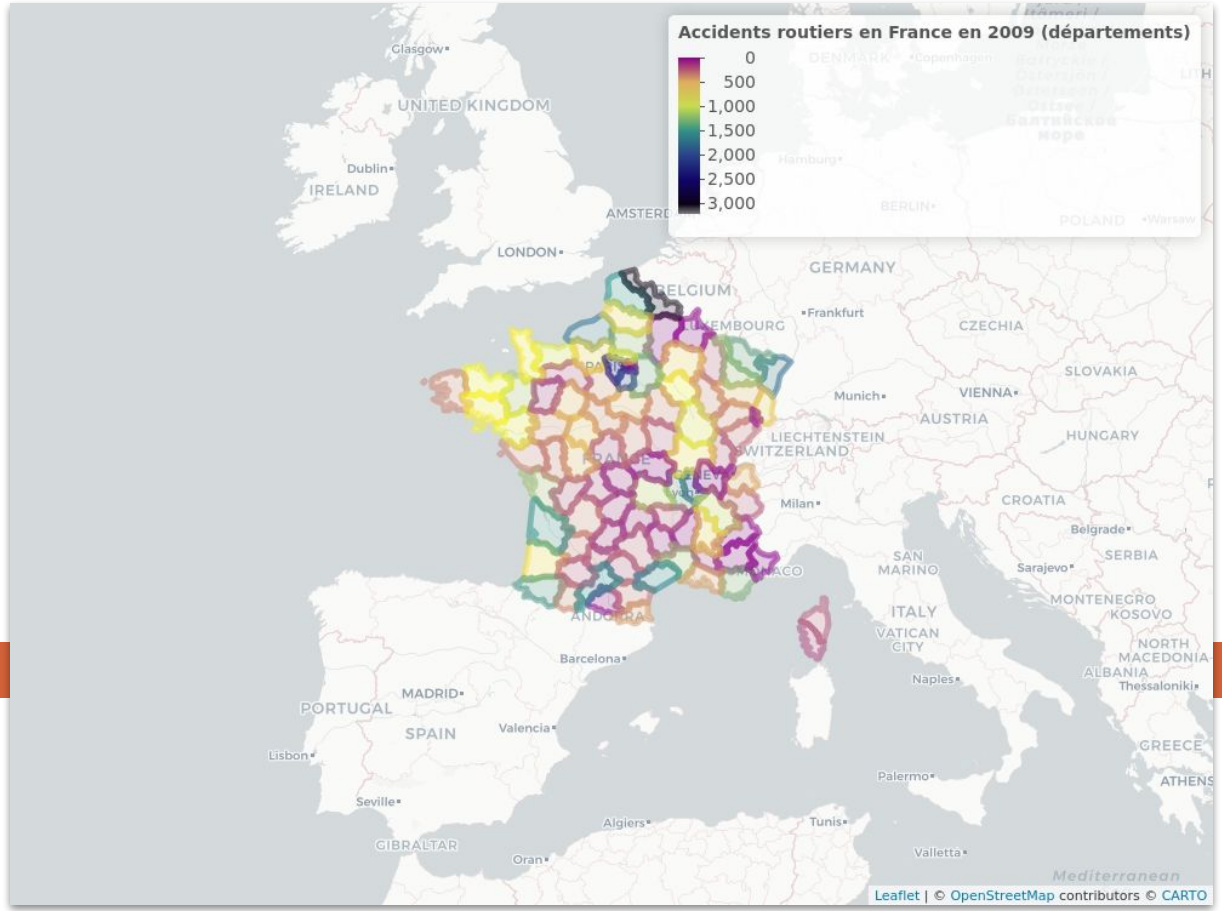
Nombre d'accidents / région



Précisions supplémentaires

- Difficultés sur l'ajout des DOM-TOM (code de région à 3 chiffres
- lien avec les régions entre le code insee et le département
- Données datés => difficultés pour trouver des cartes numériques < 2014

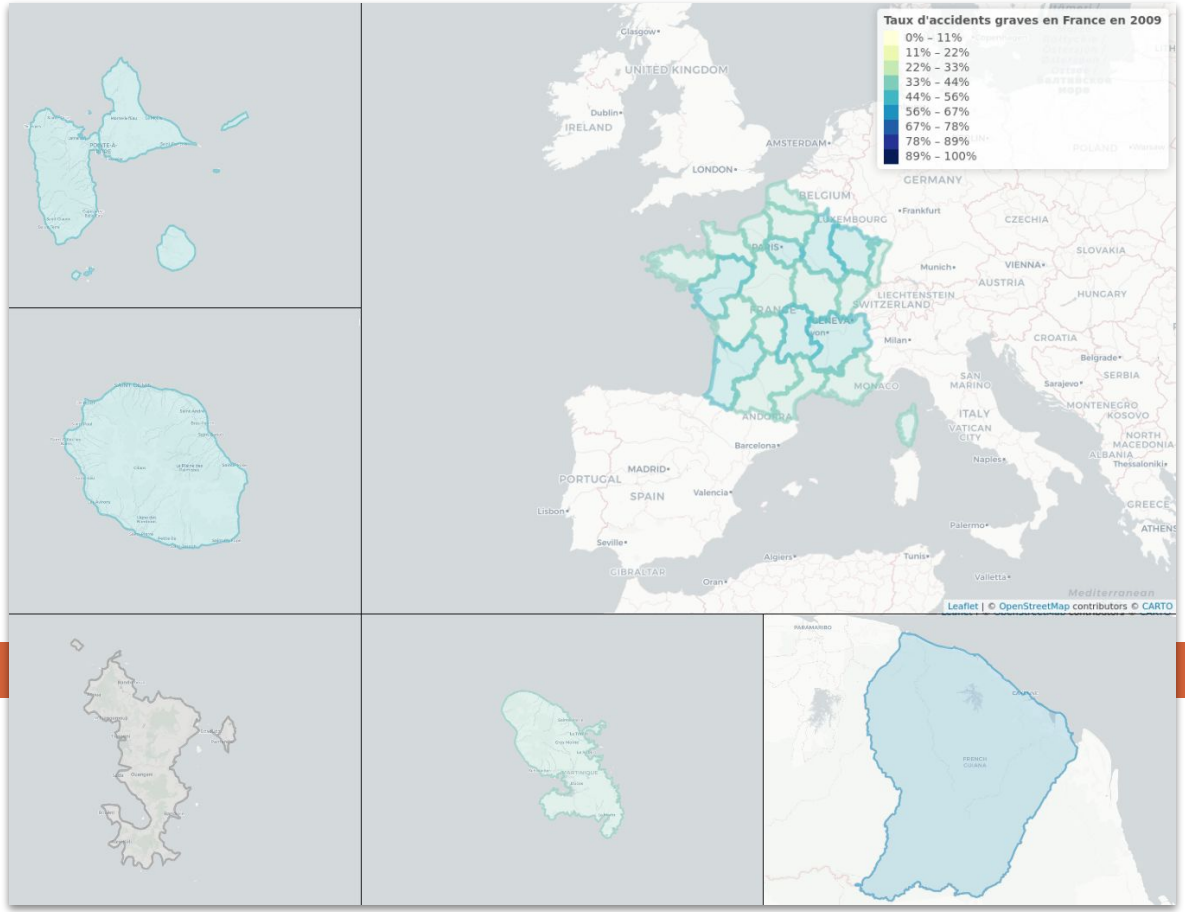
Nombre d'accidents / département



Précisions supplémentaires

- Données très hétérogènes
- Les départements à 0 accidents représentent ceux qui n'avaient pas de données enregistrés

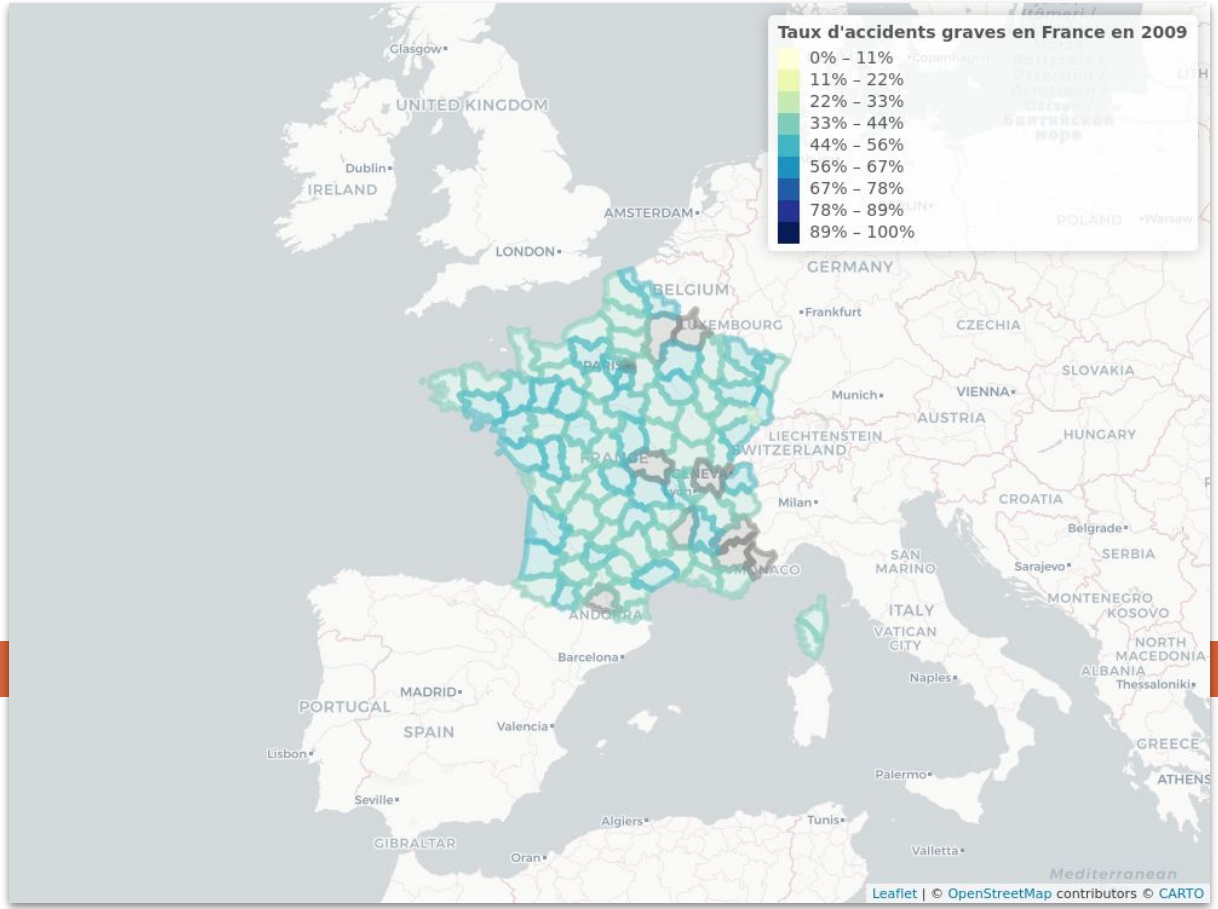
Taux d'accidents graves / région



Précisions supplémentaires

- Mayotte n'avait pas de données enregistrés (gris)
- Résultats homogènes (~40% des accidents)

Taux d'accidents graves / département



Précisions supplémentaires

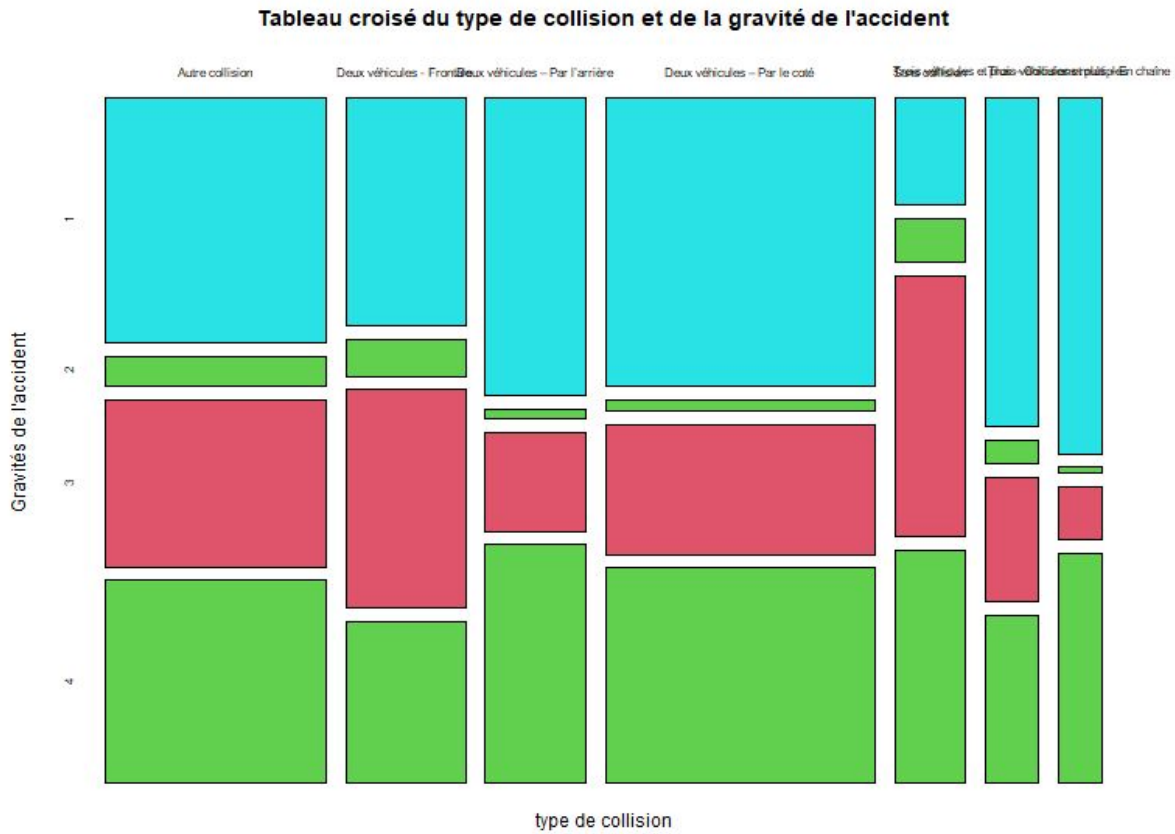
- Certains département n'avaient pas d'accidents enregistrés (gris)
- Résultats homogènes (~40% des accidents)

VARIABLES QUALITATIVES

04

Test d'indépendance X^2

- Tableau de contingence
- Statistique du khi-2
- Degré de liberté
- P-value
- Validation d'une Hypothèse(Nulle ou Alternative)



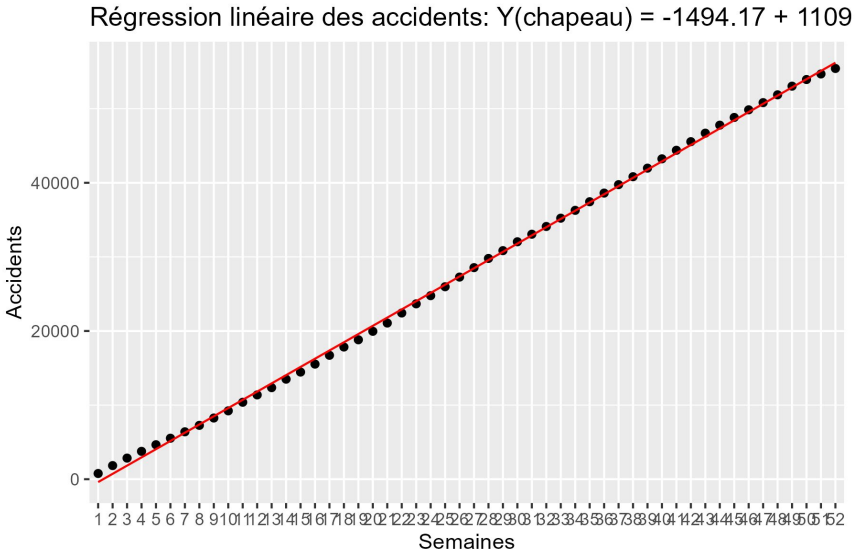
MOSAICPLOT

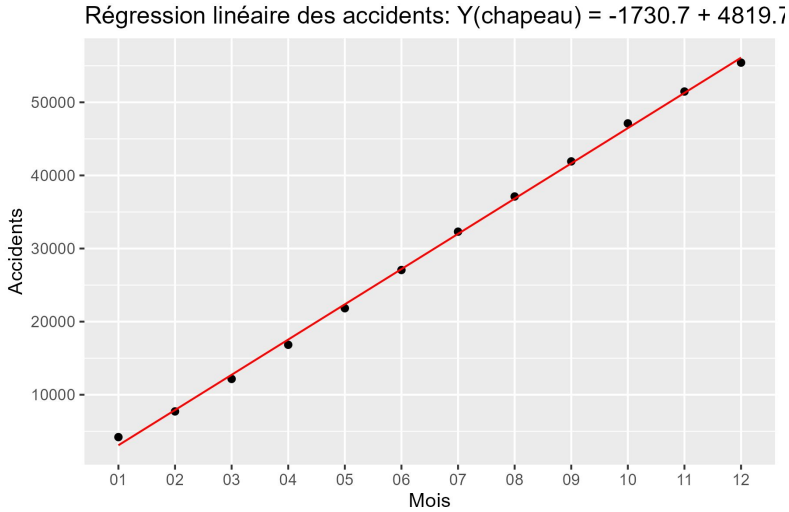
X-squared = 3563.8, df = 18, p-value < 2.2e-16

RÉGRESSIONS LINÉAIRES

05

La regression





	Evolution du nombre d’accident par mois	Evolution du nombre d’accident par semaine
Coefficient de corrélation	0.9994607	0.9995738
Intervalle de confiance à 95%	[4708.208 ;4931.3653]	[1100.359 ;1118.768]
R2(coefficient de determination)	0.9989217	0.9991479
R2 ajusté	0.9988138	0.9991308

Erreur standard des résidus

```
Call:
lm(formula = CumFreq ~ as.numeric(Var1), data = week_accident)

Residuals:
   Min     1Q Median     3Q    Max
-776.1 -382.5  103.5  262.8 1156.6

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1494.165    139.561  -10.71 1.54e-14 ***
as.numeric(Var1)  1109.564      4.583  242.13  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 496 on 50 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 5.863e+04 on 1 and 50 DF,  p-value: < 2.2e-16
```

```
lm(formula = CumFreq ~ as.numeric(Var1), data = mois_accident)

Residuals:
    Min      1Q  Median      3Q     Max
-715.45 -543.59   31.01  291.85 1116.91

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1730.70     368.56  -4.696 0.000847 ***
as.numeric(Var1)  4819.79      50.08  96.247 3.59e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 598.8 on 10 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9988
F-statistic:  9264 on 1 and 10 DF,  p-value: 3.59e-16
```

CONCLUSION

Questions?