ECOLE SUPÉRIEURE D'INGÉNIERIE EN SCIENCES APPLIQUÉES

2022

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

data.gouv.fr

REPORTING

FRENCH VEHICLES CRASHES OF 2005-2019

CREATED BY:
BENHAKKI OTHMANE

Contents

INTRODUCTION

Project Content

The following project utilize a large dataset of car crashes *(about 4.2 million records)* outsourced from the French government official website for data accuracy and legal purposes.

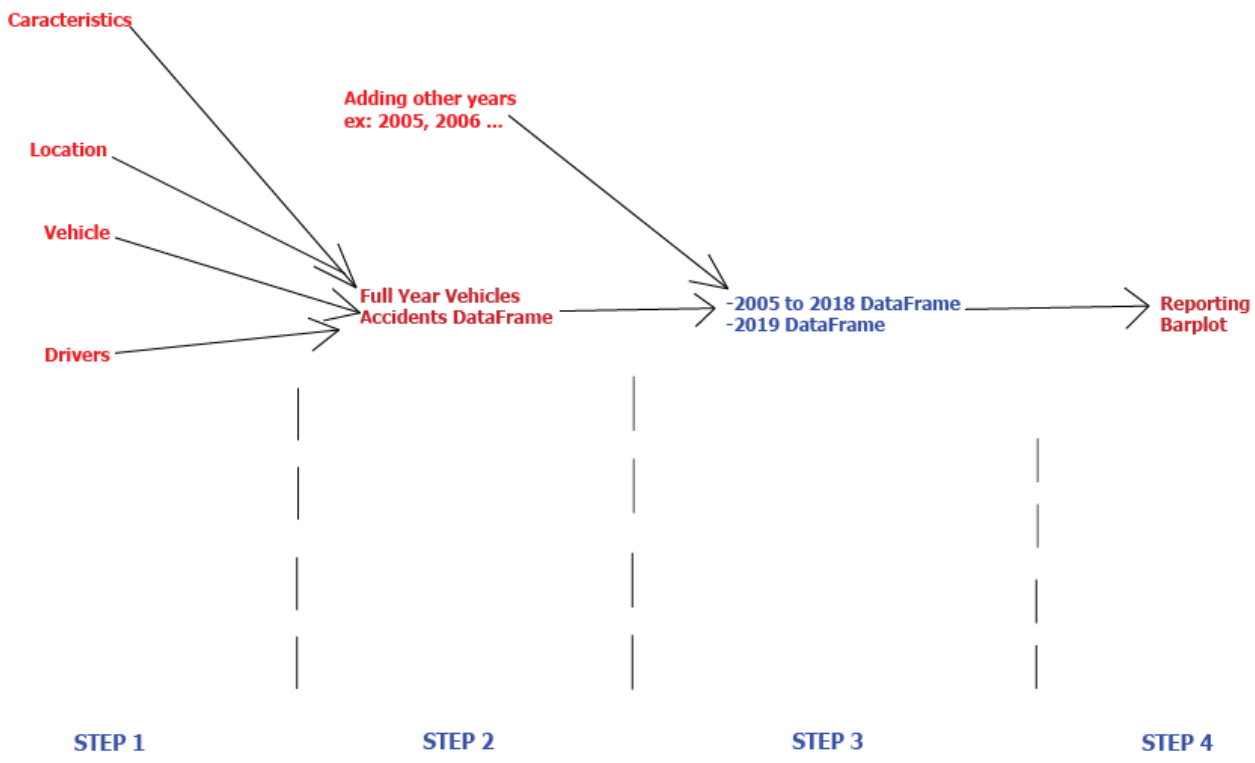
In this project, we will understand how we managed to unify and transform 56 database in a single dataset. Then the steps followed to visualize useful information in both shapes text and graphs.

Before, we start detailing any further step we must clarify the project main study goal. The objective is to answer the impact of multiple factors on accidents and compare the results with two time stamps 2005 to 2018 as reference and 2019 as a sample.

In addition, the questions are related to severity, vehicle category, road type and drivers sexe.

Project Presentation

This project followed four main steps in order to achieve the given study goal.



We have every year record separately registered on dispersed files. The shared column between all four datasets of each year is Num_Acc that represents our primary key, some important variables used:

- Grav: refers to the severity, it holds four different levels.
- Sexe: One for male and two for female.
- Num_Acc: primary key represents a unique id for each record.
- Catv: vehicle category.
- Catr : describe the road type either highway, county highway…etc.

Right now, we have a global vision of what our dataset should carry as crucial information, we will apply several modifications using R Studio. This tool is subject to be discussed in the next section.

The source of all datasets are from an official trusted governmental website:

The first step is to aggregate each year by the id. Then, the first dataset will regroup the following years (from 2005 to 2018) in a single dataset using rbind, and the second dataset will represent 2019 records only.

Now the challenge is to unify all datasets in one considering the large amount of files and the difference in delimitation between values in each table.

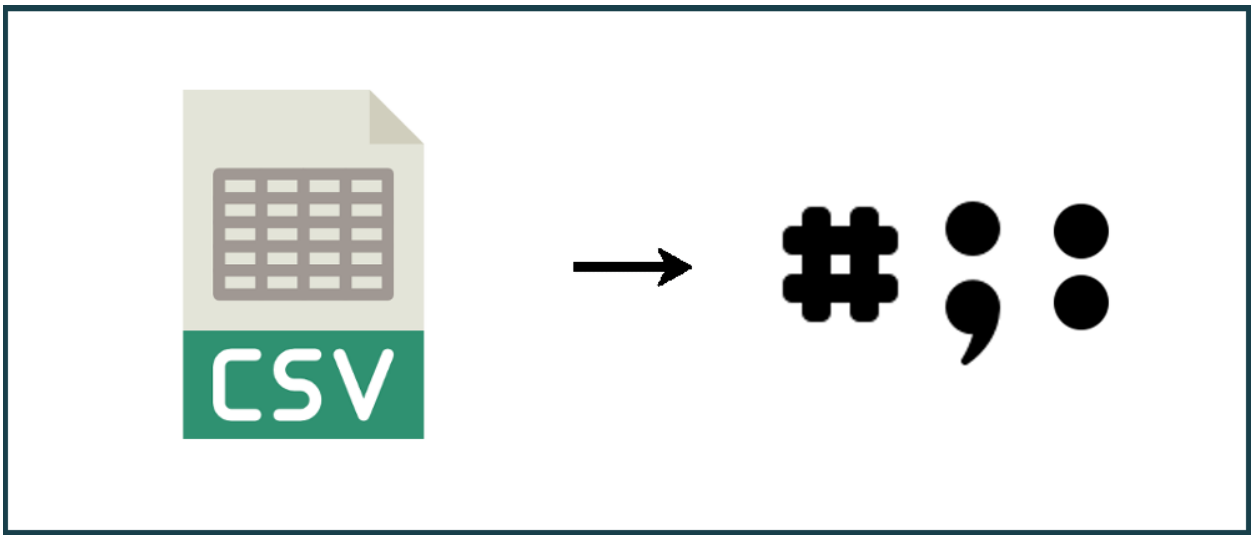
Tools Used

R Language

R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface. R is available across widely used platforms like Windows, Linux, and macOS. Also, the R programming language is the latest cutting-edge tool.

R was used in generating crucial graphs in the reporting phase, using barplot function we managed to answer all five questions and extract some relations and new factors (subject to be discussed later).

Data Presentation



The data used comes in different delimiters:

- Space: separator used mainly in header.
- Period: separator used also in header.
- Semi column: delimiter used in table content to separate columns.

<u>Why discussing this?</u>

Due to reading issues, the CSV files reading phase might generate errors. To prevent it, we need to know each file delimiters. In order to specify it in the function of read.csv or read.csv2.

Data Manipulation

Preprocessing Stage

Exception:

The dataset of 2009 ***cannot be cleaned or used anyhow***. This is due to data formatting, no separators were used which makes it impossible to identify columns values.

1	Num_Accanmoisjourhrmnlumaggintatmcolcomadrgpslatlongdep
2	200900000001913020303111667RTE DE GUï¿½MENï¿½M00440
3	200900000002911730031113 17LE BOIS JOLIM00440
4	200900000003912964531113 52M00440
5	200900000004914615311964 4LE BECOM00440
6	200900000005911615001111736MOULIN DE LA GARENNEM00440
7	200900000006912817001112756SOLFERINOM00440
8	200900000007915114511127224M00440
9	200900000008912919003111670 34 RUE DE LA HAUTE CARIZM4-440
10	200900000009911313451215126RD 723 ROUTE DE PARISM00440
11	200900000010911464531196217LA MADELEINEM00440
12	200900000011912314001261115 4RD PT ANCIENS COMBATTANTM4-440
13	200900000012913119003111714 5LA BROUINIEREM00440
14	200900000013912074531116114LA PAQUELAISM00440
15	200900000014912073052136186 17 RUE DE L'HOTEL DE VILM00440
16	200900000015912374522196471 21 BD DE LA LIBERATIONM4-440
17	200900000016911484521151118M00440
18	200900000017912418001111356VILHOUINM00440
19	200900000018911315311169 9LA TOUCHEM00440
20	200900000019913011001111113 1LE MARAIS MAINGUYM00440
21	200900000020911305211211 0AVENUE CLAUDE VELLEFAUX750
22	200900000021911818595221310 8RUE DE L ARCADE750
23	20090000(BOULEVARD MORLAND750
24	20090000(RUE DE TURBIGO750
25	20090000(QUAI BRANLY750
26	20090000(AVENUE DE FRANCE750
27	200900000026911519055231610 5RUE SAINT JACQUES750
28	20090000(BOULEVARD MACDONALD750
29	20090000(PLACE DE CLICHY750

Data Cleaning Stage

The first step in this phase is to remove unnecessary columns from the dataset. This means that we are keeping only valuable columns for this study.

The second step is to remove rows with Nans/Unknown/Blanks/Illogical or ones that can make the result less trusty. To do that, the following function 'na.omit' is used in order to do this task successfully.

```
# Removing Na and NaN Values
year.2005.to.2018.completed<- na.omit(year.2005.to.2018)
View(year.2005.to.2018.completed)

year.2019.completed<- na.omit(year.2019)
View(year.2019.completed)
```

<u>Size before cleaning:</u>

df2 × | Projet Enquete.R × | Preprocessing Agregation Combine By ... × | year.2005.to.2018.completed × | year.2005.to.2018 ×

Filter | Cols: « < 1 - 50 > »

	Num_Acc	an	mois	jour	hrmn	lum	agg	int	atm	col	com	adr	gps
2247872	2.01e+11	10	12	10	1630	1	2	1	1	6	12	227, BD JEAN JAURES	
2247873	2.01e+11	10	12	10	1630	1	2	1	1	6	12	227, BD JEAN JAURES	
2247874	2.01e+11	10	12	12	2205	5	2	1	1	6	12	95, PIERRE GRENIER (AVEN	
2247875	2.01e+11	10	12	12	2205	5	2	1	1	6	12	95, PIERRE GRENIER (AVEN	
2247876	2.01e+11	10	12	13	1225	1	2	1	1	6	12	93, REPUBLIQUE (BOULEVAR	
2247877	2.01e+11	10	12	13	1225	1	2	1	1	6	12	93, REPUBLIQUE (BOULEVAR	
2247878	2.01e+11	10	12	13	1225	1	2	1	1	6	12	93, REPUBLIQUE (BOULEVAR	
2247879	2.01e+11	10	12	13	1225	1	2	1	1	6	12	93, REPUBLIQUE (BOULEVAR	
2247880	2.01e+11	10	12	14	1300	1	2	1	1	3	12	284, JEAN JAURES (BOULEV	
2247881	2.01e+11	10	12	14	1300	1	2	1	1	3	12	284, JEAN JAURES (BOULEV	
2247882	2.01e+11	10	12	14	1300	1	2	1	1	3	12	284, JEAN JAURES (BOULEV	
2247883	2.01e+11	10	12	14	1300	1	2	1	1	3	12	284, JEAN JAURES (BOULEV	
2247884	2.01e+11	10	12	16	2115	5	2	2	3	2	12	AV E.VAILLANT/R THIERS	
2247885	2.01e+11	10	12	16	2115	5	2	2	3	2	12	AV E.VAILLANT/R THIERS	
2247886	2.01e+11	10	12	16	2115	5	2	2	3	2	12	AV E.VAILLANT/R THIERS	
2247887	2.01e+11	10	12	16	2115	5	2	2	3	2	12	AV E.VAILLANT/R THIERS	
2247888	2.01e+11	10	12	17	1435	1	2	1	1	3	12	R LES ENFANTS DU PARADIS	
2247889	2.01e+11	10	12	17	1435	1	2	1	1	3	12	R LES ENFANTS DU PARADIS	

Showing 2,247,872 to 2,247,889 of 3,999,474 entries, 52 total columns

Size after cleaning:

df2 | Projet Enquete.R | Preprocessing Agregation Combine By ... | year.2005.to.2018.completed | year.2005.to.2018

Filter | Cols: « < 1 - 50 > »

	Num_Acc	an	mois	jour	hrmn	lum	agg	int	atm	col	com	adr	gps
12664	2.018e+11	18	4	4	1730	1	2	1	1	3	287	19 rue du 4ème RSM	M
12665	2.018e+11	18	4	4	1730	1	2	1	1	3	287	19 rue du 4ème RSM	M
12666	2.018e+11	18	4	4	1730	1	2	1	1	3	287	19 rue du 4ème RSM	M
12667	2.018e+11	18	4	4	1730	1	2	1	1	3	287	19 rue du 4ème RSM	M
12805	2.018e+11	18	3	29	1427	1	2	1	1	1	89	1 Grande rue	M
12806	2.018e+11	18	3	29	1427	1	2	1	1	1	89	1 Grande rue	M
12807	2.018e+11	18	3	29	1427	1	2	1	1	1	89	1 Grande rue	M
12808	2.018e+11	18	3	29	1427	1	2	1	1	1	89	1 Grande rue	M
12809	2.018e+11	18	3	29	1427	1	2	1	1	1	89	1 Grande rue	M
12810	2.018e+11	18	3	29	1427	1	2	1	1	1	89	1 Grande rue	M
12811	2.018e+11	18	3	29	1427	1	2	1	1	1	89	1 Grande rue	M
12812	2.018e+11	18	3	29	1427	1	2	1	1	1	89	1 Grande rue	M
18330	2.018e+11	18	1	16	750	2	2	1	1	6	61	CHAMPSAUR (PLACE DU)	M
18345	2.018e+11	18	1	26	1135	1	2	1	1	3	61	CAPITAINE DE BRESSON (R	M
18346	2.018e+11	18	1	26	1135	1	2	1	1	3	61	CAPITAINE DE BRESSON (R	M
18347	2.018e+11	18	1	26	1135	1	2	1	1	3	61	CAPITAINE DE BRESSON (R	M
18348	2.018e+11	18	1	26	1135	1	2	1	1	3	61	CAPITAINE DE BRESSON (R	M
49584	2.018e+11	18	8	1	1030	1	1	1	1	7	96	CHATEAU GAILLARD	M

Showing 1 to 18 of 438,919 entries, 52 total columns

Now we must have a full ready dataset for analysis use and exploration. This step guarantee the performance of this dataset as well as the prevention of misleading results.

Data Transformation

R Studio

Merging data is the objective of this phase, thanks to R powerful functions of data manipulation and mapping, we were able to merge data in a single data frame, after of course importing all datasets to R.

Example of one year merge 2008:

```
#-----------------------------Year 2008-----------------------------------------------

caracteristiques.2008<-read.csv(file="caracteristiques_2008.csv")
lieux.2008<-read.csv(file="lieux_2008.csv")
usagers.2008<-read.csv(file="usagers_2008.csv")
vehicules.2008<-read.csv(file="vehicules_2008.csv")

#View(caracteristiques.2008)
#View(lieux.2008)
#View(usagers.2008)
#View(vehicules.2008)

year.2008.Two.Tables<-merge(x = caracteristiques.2008, y = lieux.2008, by = "Num_Acc", all = TRUE)
year.2008.Three.Tables<-merge(x = year.2008.Two.Tables, y = usagers.2008, by = "Num_Acc", all = TRUE)
year.2008<-merge(x = year.2008.Three.Tables, y = vehicules.2008, by = "Num_Acc", all = TRUE)

view(year.2008)
```

Results:

x | TP2 Analyse de donnée Fromage.R | df2 | Projet Enquete.R | Preprocessing Agregation Combine By ... | Preproc.R | year.2008 | Stats.R

Filter | Cols: « < 1 - 50 > »

	Num_Acc	an	mois	jour	hrmn	lum	agg	int	atm	col	com	adr	gps	lat	long	dep	catr	voie
1	2.008e+11	8	1	22	1015	1	2	1	1	6	424	RUE SAINT JULIEN	M	4515400	-82600	330	9	
2	2.008e+11	8	1	22	1015	1	2	1	1	6	424	RUE SAINT JULIEN	M	4515400	-82600	330	9	
3	2.008e+11	8	1	10	815	1	1	1	1	7	348		M	0	0	330	3	
4	2.008e+11	8	1	16	800	2	2	1	1	3	350	RD 231 Rue de l'hôpital	M	4478200	-12500	330	3	
5	2.008e+11	8	1	16	800	2	2	1	1	3	350	RD 231 Rue de l'hôpital	M	4478200	-12500	330	3	
6	2.008e+11	8	1	16	800	2	2	1	1	3	350	RD 231 Rue de l'hôpital	M	4478200	-12500	330	3	
7	2.008e+11	8	1	16	800	2	2	1	1	3	350	RD 231 Rue de l'hôpital	M	4478200	-12500	330	3	
8	2.008e+11	8	1	20	1100	1	1	1	1	2	394		M	4491400	-19900	330	3	
9	2.008e+11	8	1	20	1100	1	1	1	1	2	394		M	4491400	-19900	330	3	
10	2.008e+11	8	1	20	1100	1	1	1	1	2	394		M	4491400	-19900	330	3	
11	2.008e+11	8	1	20	1100	1	1	1	1	2	394		M	4491400	-19900	330	3	
12	2.008e+11	8	1	25	645	5	2	1	1	3	394	La Moleone	M	4490000	-15500	330	3	
13	2.008e+11	8	1	25	645	5	2	1	1	3	394	La Moleone	M	4490000	-15500	330	3	
14	2.008e+11	8	1	25	645	5	2	1	1	3	394	La Moleone	M	4490000	-15500	330	3	
15	2.008e+11	8	1	25	645	5	2	1	1	3	394	La Moleone	M	4490000	-15500	330	3	
16	2.008e+11	8	1	3	1000	1	1	1	8	3	14		M	0	0	330	3	
17	2.008e+11	8	1	3	1000	1	1	1	8	3	14		M	0	0	330	3	
18	2.008e+11	8	1	3	1000	1	1	1	8	3	14		M	0	0	330	3	

Showing 1 to 18 of 322,196 entries, 52 total columns

Example of 2005 to 2018 merge:

```
year.2005.to.2018<-rbind(year.2018,year.2017,year.2016,year.2015,year.2014,year.2013,year.2012,year.2011,year.2010,year.20(
View(year.2005.to.2018)
```

Results:

TP2 Analyse de donnée Fromage.R | df2 | Projet Enquete.R | Preprocessing Agregation Combine By ... | Preproc.R | year.2005.to.2018 | Stats.R

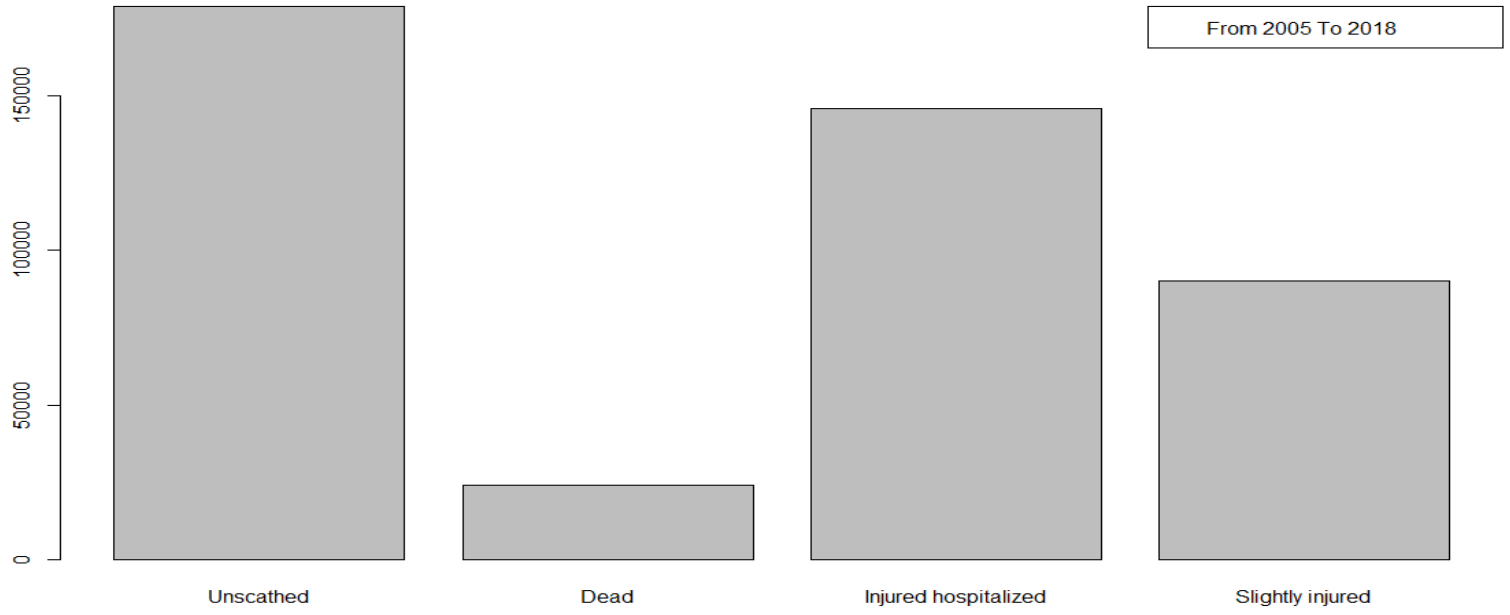
Filter | Cols: « < 1 - 50 > »

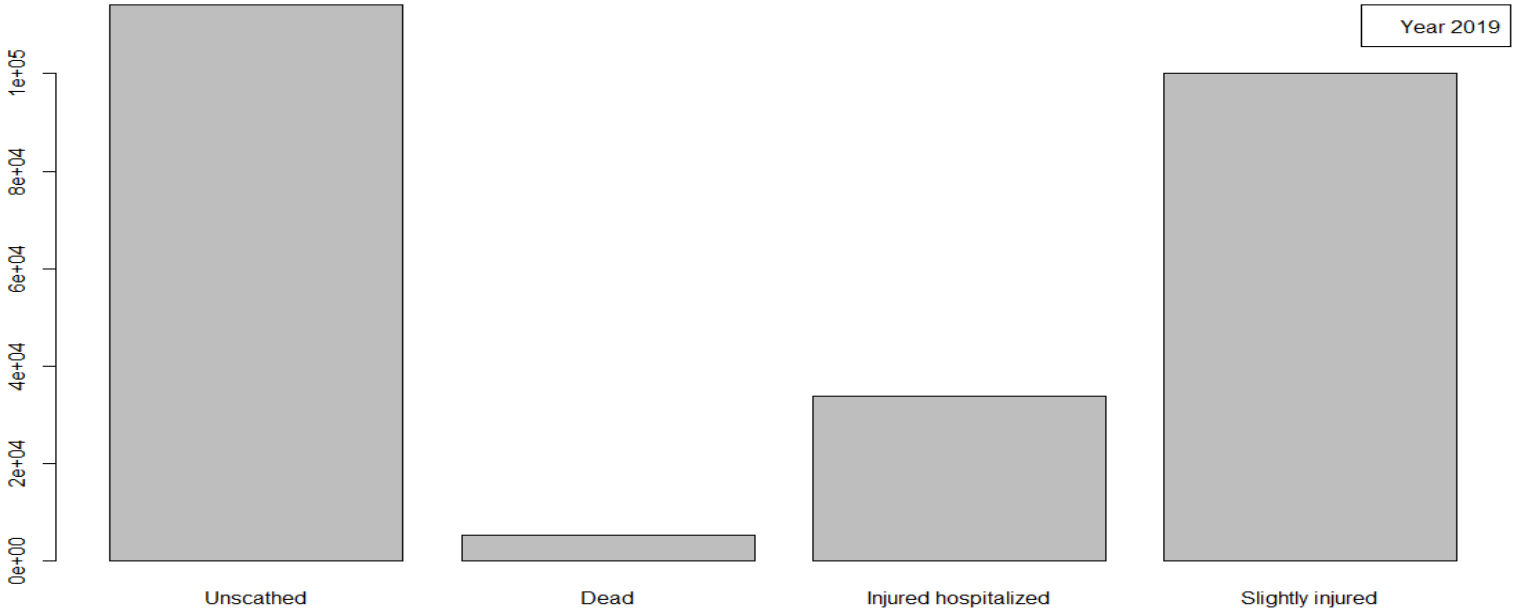
	Num_Acc	an	mois	jour	hrmn	lum	agg	int	atm	col	com	adr	gps	lat	long	dep	catr	voie
1	2.018e+11	18	1	24	1505	1	1	4	1	1	5	route des Ansereuilles	M	5055737	294992	590	3	41
2	2.018e+11	18	1	24	1505	1	1	4	1	1	5	route des Ansereuilles	M	5055737	294992	590	3	41
3	2.018e+11	18	1	24	1505	1	1	4	1	1	5	route des Ansereuilles	M	5055737	294992	590	3	41
4	2.018e+11	18	1	24	1505	1	1	4	1	1	5	route des Ansereuilles	M	5055737	294992	590	3	41
5	2.018e+11	18	2	12	1015	1	2	7	7	7	11	Place du général de Gaul	M	5052936	293151	590	4	41
6	2.018e+11	18	2	12	1015	1	2	7	7	7	11	Place du général de Gaul	M	5052936	293151	590	4	41
7	2.018e+11	18	3	4	1135	1	2	3	1	7	477	Rue nationale	M	5051243	291714	590	3	39
8	2.018e+11	18	3	4	1135	1	2	3	1	7	477	Rue nationale	M	5051243	291714	590	3	39
9	2.018e+11	18	3	4	1135	1	2	3	1	7	477	Rue nationale	M	5051243	291714	590	3	39
10	2.018e+11	18	3	4	1135	1	2	3	1	7	477	Rue nationale	M	5051243	291714	590	3	39
11	2.018e+11	18	3	4	1135	1	2	3	1	7	477	Rue nationale	M	5051243	291714	590	3	39
12	2.018e+11	18	3	4	1135	1	2	3	1	7	477	Rue nationale	M	5051243	291714	590	3	39
13	2.018e+11	18	5	5	1735	1	2	1	7	3	52	30 rue Jules Guesde	M	5051974	289123	590	3	39
14	2.018e+11	18	5	5	1735	1	2	1	7	3	52	30 rue Jules Guesde	M	5051974	289123	590	3	39
15	2.018e+11	18	5	5	1735	1	2	1	7	3	52	30 rue Jules Guesde	M	5051974	289123	590	3	39
16	2.018e+11	18	5	5	1735	1	2	1	7	3	52	30 rue Jules Guesde	M	5051974	289123	590	3	39
17	2.018e+11	18	6	26	1605	1	2	1	1	3	477	72 rue Victor Hugo	M	5051607	290605	590	4	
18	2.018e+11	18	6	26	1605	1	2	1	1	3	477	72 rue Victor Hugo	M	5051607	290605	590	4	

Showing 1 to 18 of 3,999,474 entries, 52 total columns

Data Reporting

Deadly Crashes Rate





The following graph shows the rate of death, 25 thousand person during the years of 2005 and 2018. However, in 2019 this study show that the death rate decreased and reached only 5 thousand person. To conclude, the rate heavily decreased by -20%.

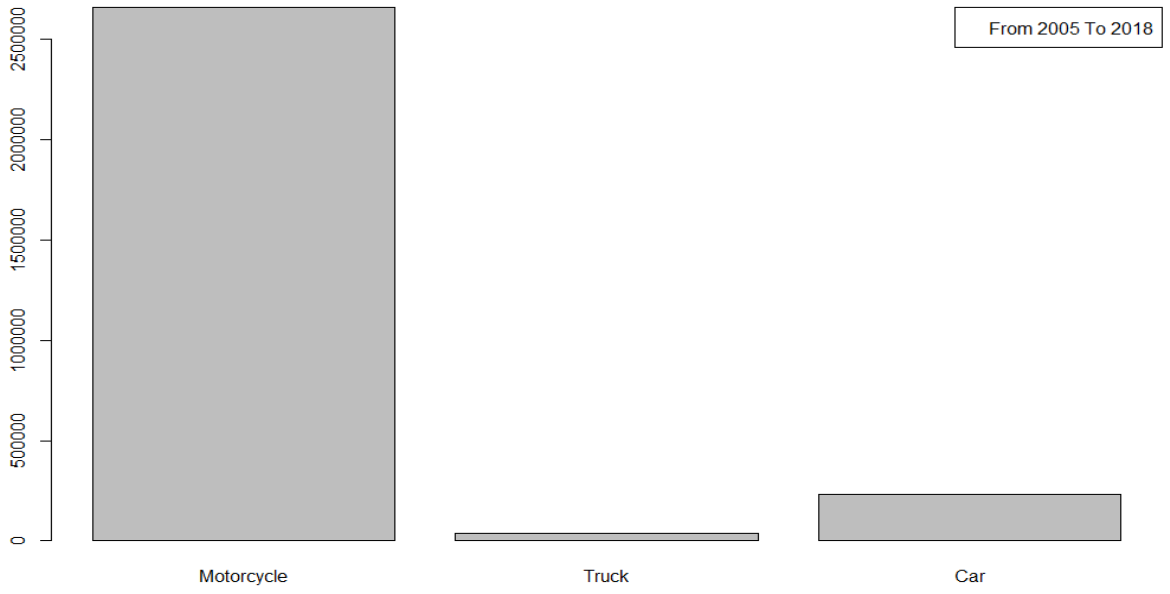
Injured Hospitalized Crashes Rate

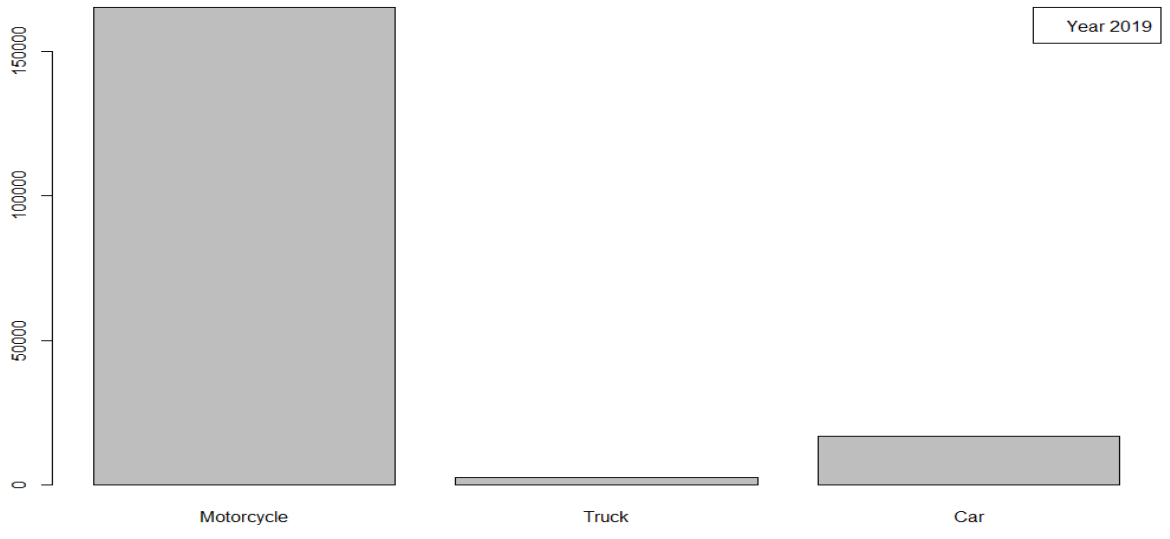
The same graph shows the rate of injured hospitalized, 150 thousand person during the years of 2005 and 2018. However, in 2019 this study shows that this rate decreased and reached only 35 thousand person. To conclude, the rate heavily decreased by -23%.

Hypothesis:

This decrease can be related to the precautions and the new driving laws that where implied by the French government, or might be related to the developed safety technologies used in newer vehicles.

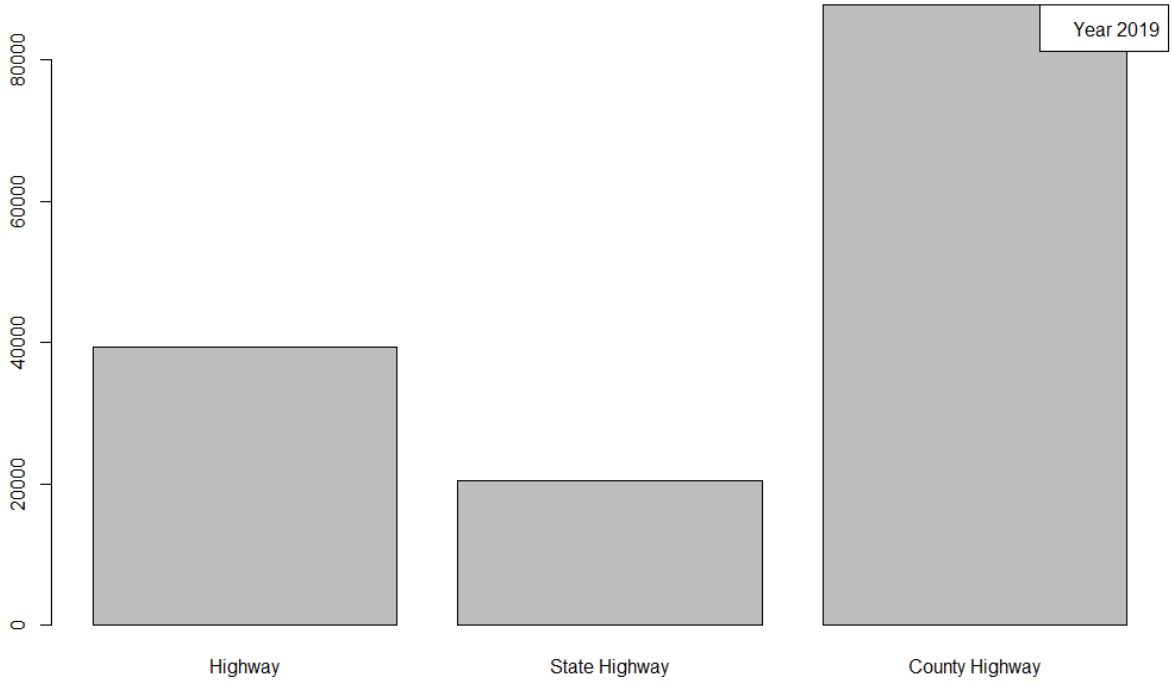
Crash Rate By Vehicle Category

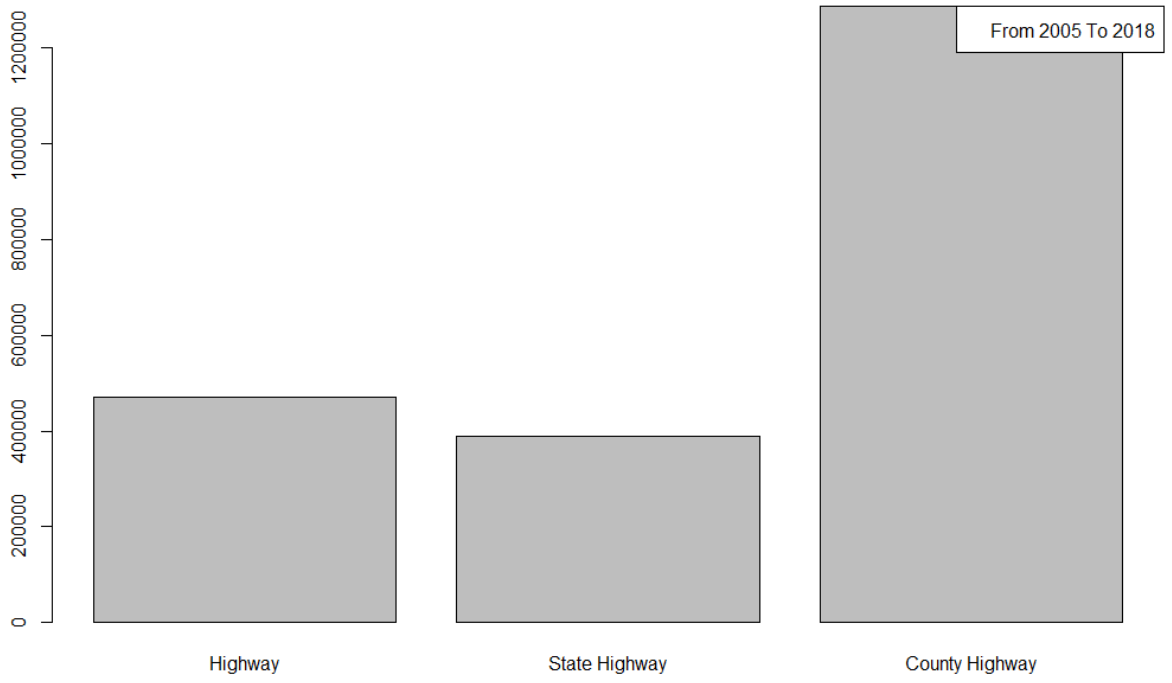




This statistical report ensures that rate of two wheelers is significantly decreased since 2005 by 96.4%. However, this study shows lower rate for truck category up to -10% for 2019. Also, the car category had known ~ -10% in total accidents rate.

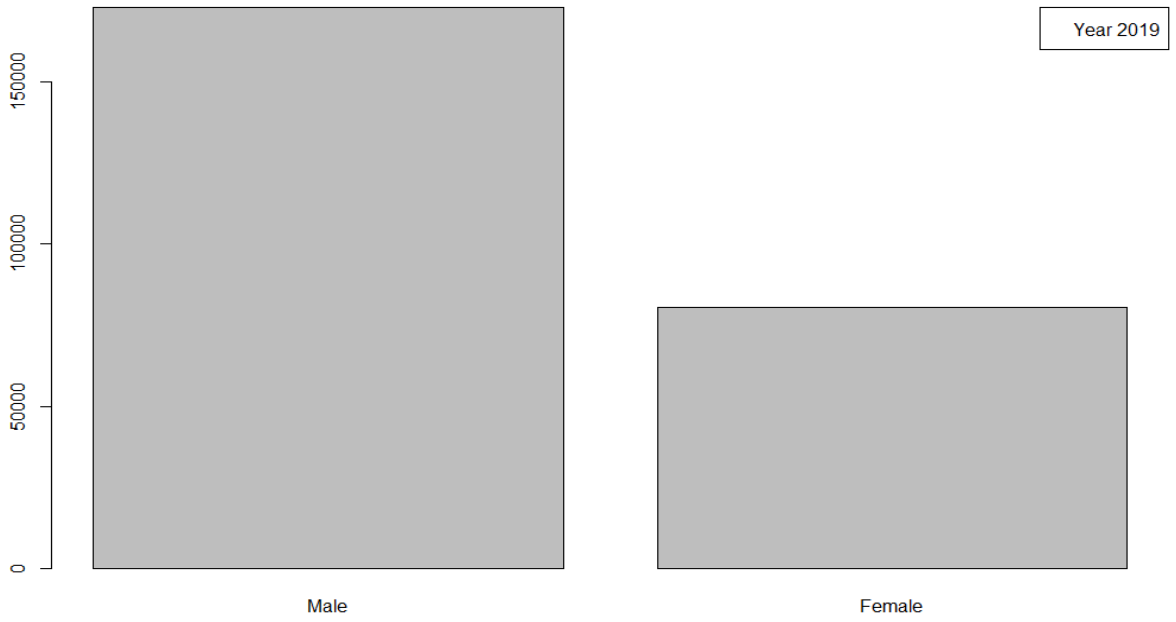
Crash Rate By Road Type

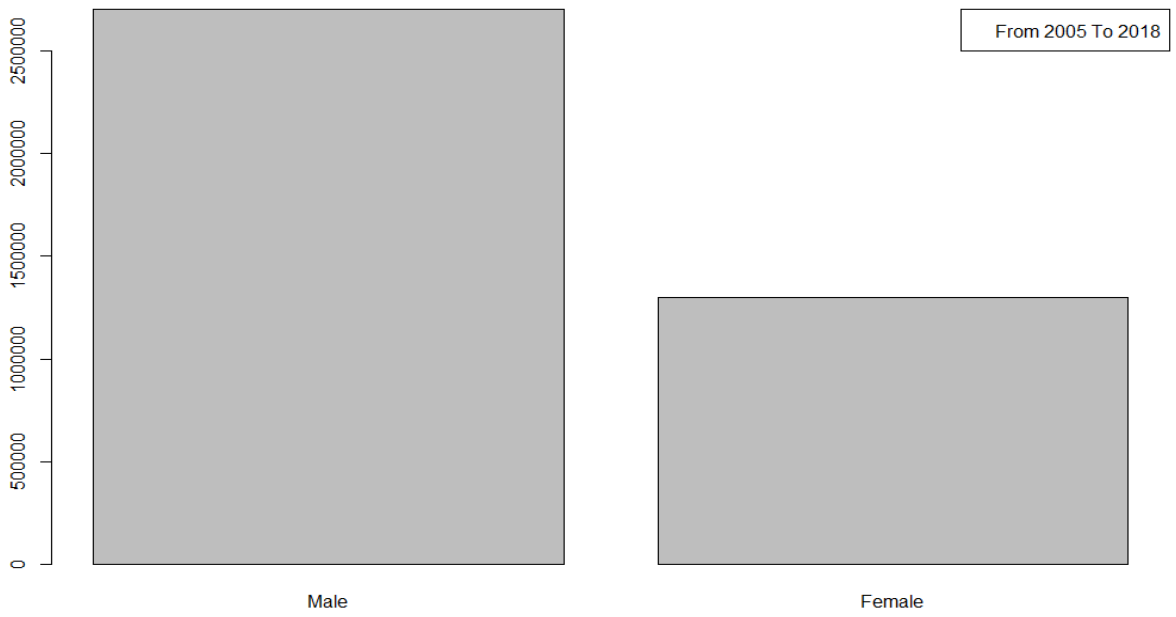




As observed, we can clearly ensure that highways are not the most dangerous roads. Otherwise, county highways are considered very dangerous and have higher risk of collision. However, this statistics can be empowered by the fact that highways are easier to drive in, due to the factor of having straight roads and wide surfaces.

Crash Rate By Gender





Gender might not be considered as a factor, but this statistics changes this statement. Due to the fact that males have twice accident rate than females, so it might be added to causes or factors leading to collisions.

Conclusion

This study helped to extract three crucial criteria's that exists on every accident. First, two wheelers have higher rate of accidents than other vehicle categories. Second, a crash is more likely to occur 2X times more on county highways. Third, gender is considered a solid crash factor due to higher rate of collisions for males compared to females. To conclude, those three components are the factors of having a high chance to a collision on public French roads.

TWO WHEELER + COUNTY HIGHWAY + MALE = SEVERE ACCIDENTS

9 781234 567897 12345