Road accidents in France

DataScientest - Aug25 Bootcamp: Data Analyst

Oscar CANIVET

Naomi Thandi KLINGBEIL

Markus NORDMANN

October 28th, 2025

Table of contents

Part A: exploration, data visualization and data pre-processing report

1. Introduction to the project

1.1 Context

This project focuses on the analysis of road traffic accidents in France between 2019 and 2023. The data comes from the official governmental database that has recorded over thousands of road accidents occurring every year in France (Metropolitan and overseas). The data allows us to see the many factors that have an impact on the number of people in road accidents but also the severity of these road accidents. Understanding these factors and their potential influence are essential for developing effective prevention strategies and improving infrastructure design to lower severe outcomes.

Our analysis primarily focuses on how accident severity may vary according to several factors that we chose out of the many available: temporal (month, hour, weekday, and year), user profile (age), location (department), and road conditions (speed limit). We placed a special emphasis on vulnerable road users, in this case cyclists and e-bike users whom we believe are exposed to a greater risk in recent years.

Using the concepts studied throughout our bootcamp, such as data visualization and statistical analysis, this first report presents the methodological steps we followed to prepare and analyze the data, along with our reasoning, visualizations (graphs and tables), and statistical results. The second part of this report will present our interactive Power BI dashboard, summarizing the building process and the interpretation of our results.

While this does not relate to any team members current professional role, it remains a vital training exercise to gain handson experience with large database analysis, visualization design, and storytelling through data that may be used as a portfolio piece for future job applications. For the reader, this report is meant to illustrate the steps we followed to clean, manipulate, and analyze the data, and how we transformed it into a dynamic, interactive Power BI dashboard.

At the same time, it aims to provide a better understanding of road safety dynamics in France and to identify which factors (in the ones we have chosen), have a measurable impact on injury severity.

1.2 Objectives

The main objectives of this project are:

- Analyze and understand the different factors that influence road accidents in France (2019-2023).
- Conduct a data audit (and use Python) to evaluate data quality, identify missing values, and help us choose which columns to keep or drop.
- Clean, standardize, and rename columns for consistency, and create new columns (i.e. create age column using columns year of birth and year of accident).
- Using seaborn and matplotlib to create visualizations (graphs) of our chosen factors in terms of number of people in accidents and injury severity.
- Perform statistical analyses (i.e. Chi-square tests, Cramer's V, Logistic Regression) to test relationships between severity and chosen factors.
- Develop an interactive Power BI dashboard of this data for further (and improved) visualization.
- Summarize findings/results and share conclusions drawn on possible recommendations to improve road-safety from findings.

1.3 Team expertise

Thandi has taught statistics at a university in the USA and enjoys data modeling. She loves running and cycling and appreciates road safety when she's out there.

Oscar previously worked as a junior revenue manager for a hotel located in Paris. He discovered there an interest in learning tools such as Microsoft Excel and wants to learn more through this bootcamp and project.

Markus works in an auditing firm and has extensive experience with financial data. In his free time, he is a passionate cyclist, which motivates his specific interest in analyzing bicycle users as accident participants in this project.

2. Understanding and manipulation of data

2.1 Framework

The data used for this project is freely available from a public database available on the website www.data.gouv.fr that lists all road traffic accidents involving injuries that occurred during a specific year in mainland France, in the overseas departments (Guadeloupe, French Guiana, Martinique, Réunion and Mayotte since 2012) and in the other overseas territories (Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Martin, Wallis-et-Futuna, French Polynesia and New Caledonia; only available from 2019 in the open data.
For each accident involving bodily injury (i.e. an accident occurring on a road open to public traffic, involving at least one vehicle and resulting in at least one victim requiring treatment), information describing the accident is entered by the law enforcement unit (police, gendarmerie, etc.) that responded to the scene of the accident. These entries are collected in a form called a bodily injury accident analysis bulletin. All these forms constitute the national file of bodily injury road accidents known as the *BAAC File* administered by the National Interministerial Road Safety Observatory (ONISR).
The databases from 2005 to 2023 are annual and made up of 4 files (Characteristics – Places – Vehicles – Users) in csv format.
For this project, we consider the data from the years 2019 to 2023, which includes information on 273 226 accidents during the five years.

The question asked in this analysis was: “Does injury severity (severe = killed or hospitalized wounded vs non-severe) depend on factors like the time of day, month, weekday, year, driver characteristics (such as age), location and road characteristics (such as the speed limit on the road where the accident occurred)?”
Logistic regression, Chi-square tests, and Cramér’s V were used to examine relationships between injury severity and the different variables.
Logistic regressions identified which categories were more or less likely to result in severe outcomes. The Pseudo R^2 values were small, indicating that although patterns exist, these variables explain only a small part of the variation in injury severity.
Chi-square tests confirmed that the factors were significantly linked to severity.
Cramér’s V values, however, showed these relationships were weak in strength, suggesting that while statistically significant, they explain only a small share of overall injury severity patterns.

2.2 Relevance

Understanding which factors contribute to severe road accidents is of high societal importance, particularly as recent data from ONISR highlight an increasing share of vulnerable road users among the most severely injured. According to ONISR, “Since the pandemic, the proportion of vulnerable road users, i.e. those without a body (pedestrians, cyclists, EDPm users, motorized two-wheeler users) among those killed or seriously injured has increased. Car occupants now account for less than half of those killed (48%). The proportion of motorized two-wheeler users is increasing; they represent 23% of those killed, 32% of those seriously injured and 36% of those injured who will have after-effects 1 year after the accident, for less than 2% of motorized traffic. The proportion of cyclists and users of EDPm is increasing; they represent 8% of deaths, 21% of serious injuries and 32% of injuries who will have after-effects 1 year after the accident.” (Bilan 2024 de la sécurité routière)

In this analysis our target variable is injury severity grouped into three categories: non-severe, hospitalized wounded, and killed.

Given the observed rise in severe outcomes among cyclists, a particular focus of this project is to analyze a few factors influencing the severity of accidents involving this group.

2.3 Pre-processing and feature engineering

We first imported the aggregated files from www.data.gouv.fr found in the community resources tab, which cover the time period 2005 to 2021 for the four main tables: Caractéristiques, Usagers, Lieux, and Véhicules (Translates to: Characteristics, Users, Places, & Vehicles). However, some of the columns had inconsistent data types across years (i.e. integer in one year, float in another), and the aggregated files contained columns such as a gps column that is not in the database, but was added by the user. There is no documentation on how this column was created and the overseas departments Guadeloupe, French Guiana, Martinique, Réunion and Mayotte were added after 2012, not since 2005. The overseas territories (Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Martin, Wallis-et-Futuna, French Polynesia and New Caledonia) are only available from 2019 in the open data.

Speed limit on the road on which the accident occurred is another such variable that was added to the database from 2019 onwards. Since our variable of interest is the severity of the injuries and we assume that speed could be a factor in determining the severity of the accident, we would like to look at all the data for this variable. User_id, tracking the actual people involved in the accident was another column that was added in 2021.

Based on the above factors, we decided to instead retrieve the individual year files for 2019 – 2023 (five years for each table, giving us a total of 20 files). We added a year column and concatenated each of the five files for Characteristics, Places, Vehicles and Users.
A thorough data audit was done on all variables in the four files.
We cleaned up the different data types and column name differences (*Accident_Id* versus *Num_Acc*) in all files across the years. We checked for duplicates in the files and found that the Users file had 164 duplicates. These were dropped from the files.
We then merged the four concatenated files. *Num_Acc* is the unique identifier of the accident in the Characteristics file. In the three remaining files, *Num_Acc* is not unique since there are cases of more than one person per accident (Users file), more than one vehicle per accident (Vehicles file) and more than one place (such as two different roads with different speed limits) recorded per accident (Places file). The merged dataset is structured so that each row represents an individual involved in an accident, since our primary variable of interest is the person's injury severity. We anchored on the Users file and made every other join many-to-one with respect to the user row. Starting with the Users file (619 807 users recorded) and did a left merge with the Vehicles file on a composite key of the three columns that these two files have in common (*Num_Acc, id_vehicule* and *num_veh*). We then merged the Characteristics file on the *Num_Acc* column, which is unique, and so we had no row explosion. The Places files proved more challenging, as it has multiple rows per accident and only one column in common with the Users file (namely *Num_Acc*). Merging it as is, would lead to row explosion. Hence, we aggregated the rows in the Places file into pipe-separated lists to reduce it to one row per accident (without adding extra columns) before merging. Only different values were stored in pipe-separated lists; for identical values, one value was kept. The final merge resulted in 55 columns and 619 807 rows of data (one row per person involved in each accident) for the 273 226 accidents during the five years.

During the inspection of the injury_severity variable, we identified a small subset of records with the value -1, corresponding to "Not specified" cases. This category represents only 0.07% (419 out of 619 807) of all records. Due to its very low frequency, the impact of these records on the overall analysis is considered negligible.

To simplify the analysis and maintain a consistent categorization across the dataset, the cases with the value -1 were included in the "Non-severe" category together with the unscathed and lightly injured cases. Although this decision is not semantically perfect, it provides a cleaner structure for the analytical process and avoids unnecessary fragmentation of the data while ensuring statistical robustness.

During the inspection of the year_of_birth field, we found a subset of records coded as NaN/-1, indicating “Not specified”. This affects the later derivation of age (accident year minus year_of_birth) and the corresponding age buckets. The NaN/-1 cases represent 1.38% (8 539 out of 619 807) of all records, so their influence on aggregate results is limited.

To preserve data integrity, we kept these raw values in the dataset, but we did not display the NaN/-1 category in visualizations. For computed fields, these records are treated as Unknown in age and are excluded from age-bucket charts and KPIs. This approach avoids misleading buckets or forced imputations, keeps the underlying data reproducible, and prevents a very small set of unspecified values from adding noise to the graphs.

The speed limit (originally vma) column from the non-aggregated Places file also contained -1 values (7 971 out of 289 264, or 2.76%), similar to the year_of_birth column mentioned above. These cases were dealt with in the same way as outlined for the ages in terms of the speed buckets and for the corresponding graphs. Aggregating the rows of the Places file into pipe-separated lists before merging, meant that we had to deal with the 5 169 resulting rows where the speed limit was a pipe-separated list. If the speed limits on all roads recorded for the accident were the same, they were not aggregated, and were not a pipe-separated list. There were 5 169 cases of different speed limits, or one speed limit recorded and the other coded as -1 (not specified). In the cases where one speed limit was recorded and the other unspecified, we kept only the recorded speed limit, and dropped the pipe. In the cases where two or more different speed limits were recorded, we dropped the pipe and kept the average speed limit as the single value for the column.

2.3.1 Helper function: checking unique values

To support our data exploration and cleaning process, we defined a helper function checkcolumn().
This function returns the counts and percentages of unique values in a given column, including missing values.

It also offers two sorting options:
sort = 0 - sort results by the column index (ascending)
sort = 1 - sort results by frequency (descending, default)

This allows us to quickly inspect the distribution of values across different variables and identify potential anomalies or imbalances in the dataset.

2.3.2 Dropping columns

Based on the percentages of missing data and the usefulness of certain columns for our analysis, we decided to drop the following columns:

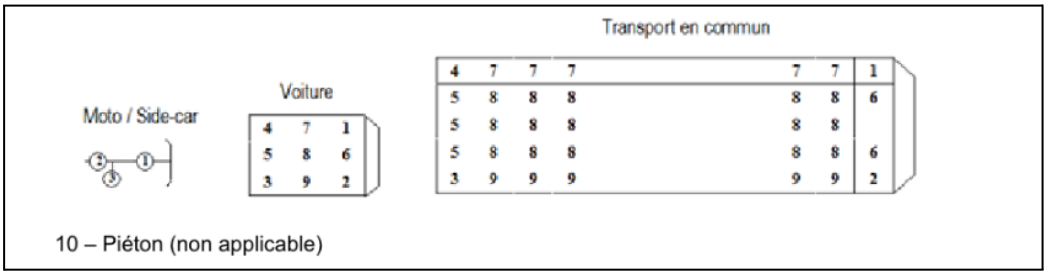
Characteristics file:

- Adr [Postal address, object, 1.46% missing data (md)]: The postal code in France is the following format "XX345", with XX being the department number (i.e. Paris, department: 75, example postal code: 75018). However, we already have more precise and easier-to-use columns (in terms of location) such as "com (Commune, object, 0% md)", "lat (Latitude, object, 0% md)", or even "dep (Department, object, 0% md). This makes the "Adr" redundant and less convenient to use for analysis therefore we dropped this column.

Vehicles file:

- Senc (Direction of travel/circulation, integer, 0% md): This column takes into consideration PK (Kilometer Point) and PR (Reference Point) that we can find on road markers and sees whether they are (0 - unknown, 1 - increasing, or 2 decreasing). This information is not relevant to our main objectives (understanding accident trends using the measurement of severity). Since it only tells us the roadway orientation and does not affect the variables we are analyzing (such as user vehicle types, road type, or weather conditions).
- Occutc (Number of passengers in public transport, float, 99.16% md): This column takes public transports (i.e. bus, trams, etc.) and gives us the number of passengers at the time of accident. However with a high percentage of missing data we dropped this column.

Users file:



- Place (Location in vehicle, integer, 0% md): Using the image above as an example, this column provides a number from -1 to 10 which communicates whereabouts in the vehicle the person in an accident was located. For our research we are interested in knowing the category of the user in the accident. Though this information can be found much simpler in the column "catu (user category, integer, 0% md)" which only uses numbers 1 for driver, 2 for passenger, and 3 for pedestrian. Therefore the column "catu" is kept and the "place" column is dropped.

- Locp (Location of pedestrian, integer, 45.19% md): This column (using numbers from 1 to 9) provides us with information such as was a pedestrian within 50 meters or not from a pedestrian crossing, did they cross with or without a traffic light, and other information such as on sidewalk, emergency lane, etc. As not all accidents in this data involve a pedestrian we have a good chunk of data missing and therefore this column is dropped.
- Actp (Action of pedestrian, object, 40.15% md): This column provides, using numbers and letters, the direction a pedestrian was moving (i.e. towards or further from vehicle involved), if they were playing/running, with an animal, and if they were getting on or off a vehicle. It has a high portion of data missing so this column is dropped.
- Etatp (Individual/Accompanied/Group Pedestrian, integer, 92.25% md): This column uses numbers 1 to 3 to specify if the pedestrian involved in the accident was: 1 for alone, 2 for accompanied, or 3 for in a group. This column is dropped due to a high percentage of missing data.
- Id_usager (User Id, string, 38.43% md): Our data starts from the year 2019, however this column first makes an appearance in the year 2021-users file thus 2019 and 2020 does not have this information. This column helps see when a specific user (represented by an Id) has been involved in one or multiple accidents. This is meant to be used as a primary key most likely for a table that we do not have access to that would provide the names and other information of the users involved in an accident. As we have a good chunk of this data missing and we do not have access to the other table we have dropped this column.

Places file:

- Voie (Road number, object, 12.38% md): Provides us with the road name or number where the accident took place. This column is directly related to two other columns that we also dropped called “v1 (numeric index of road, float, 3.71% md)” and “v2 (letter index of road, object, 91.99%)”. We are not interested in the specific road name where an accident took place as this would create too many different categories to measure. We prefer aspects we can group such as road alignment (curved left, in “S”, straight, etc..) or in which department (giving us less possibilities in terms of results compared to “Voie”).

- Pr (Reference point, object, 0% md) et PR1 (Distance from Reference point, object, 0% md): Roads in France have markers, the image above is an example. These markers provide us with our reference point, the letter

represents road category (i.e. N is national road), and the number next to the letter is used as further identification (i.e. N77 is the road that connects the city of Troyes and Auxerre). The kilometer point is underneath and helps us locate where specifically on the road we are, however in this case PR1 takes the distance we are from a reference point in terms of meter. If we wanted to measure for example how long it took for an ambulance to get to that point and compare it over time we would use the kilometer point. This is not interesting with what we are trying to measure and remains similar to the “Voie” column where it is too specific of a condition if we want to measure it with severity and not offer us the potential to identify trends since the data will most likely be spread out everywhere and it is unlikely to find common points (Especially when we have columns like department that can better provide this). For those reasons these two columns have been dropped

- Lartpc (Width of central median, float, 99.82% md): This column provides data that is too specific. It takes into account the width (in meters) of the divider/strip separating directions on a road.
- Larrout (Width of road, float, 20.21% md): This is the width of a road (in meters) that does not take into account the width of the central median, the emergency lane, or parking spaces. This information would have been extremely interesting when we take into account that we also have plenty of information regarding our road that we could relate it to (i.e. traffic regime: one way, bidirectional) though 20.21% of missing data is too high so we decided to drop it.

2.3.3 Renaming columns

This data comes from a French governmental website that provides the table with the column names in French. A good portion of these columns are in abbreviated form (i.e. The column VMA is French for “vitesse maximale autorisée”). There is a document that provides further explanation to what these columns represent. Therefore we renamed these columns to English to know what each represents and avoid confusions. The Data Audit included the column name, description, variable type, % of missing data. We decided on a schema such as the word ‘number’ being represented as ‘num_’ and always being the prefix of a column name.

We renamed the columns as outlined below:

```
"Num_Acc": "num_accident",    #  from Characteristics Table
"jour": "day",
"mois": "month",
"lum": "light_cond",
"agg": "loc_type",
"int": "intersc_type",
"atm": "weather_cond",
"col": "collision_type"

"senc": "direc_travel",              # from Vehicles Table
"catv": "category_vehicle",
"obs": "fixed_obs_hit",
"obsm": "mobile_obs_hit",
"choc": "int_impact_point",
"manv": "maneuvre"

"id_vehicule": "id_vehicle",       # from Users Table
"num_veh": "num_vehicle",
"catu": "category_user",
"grav": "injury_severity",
"sexe": "gender",
"an_nais": "year_of_birth",
"trajet": "trip_purpose",
"secu1": "safety_equipment1",
"secu2": "safety_equipment2",
"secu3": "safety_equipment3"

"catr": "category_road",          #  from Places Table
"circ": "traffic_reg",
"nbv": "num_traffic_lane",
"vosp": "reserved_lane",
"prof": "road_slope",
"plan": "road_alignmt",
"surf": "road_surface",
"situ": "accident_situ",
"vma": "speed_limit"
```

2.3.4 Creating columns

Age and age buckets

To make age-related analysis easier, we first calculated the age of each user by subtracting their year of birth from the accident year. We ensured data quality by coercing invalid entries to NaN and storing the result as an integer type (Int64).

Since plotting every single age would create too much clutter in visualizations, we grouped ages into age buckets. These buckets represent broader ranges (i.e. 18–24, 25–34, 85 and older) and provide a clearer overview of demographic patterns in the dataset. These age buckets are the same ones found on the ONISR website.

Date, Datetime, hour and weekday

The dataset originally provided separate fields for day, month, year, and time.
To perform meaningful temporal analysis, we combined these into new columns:

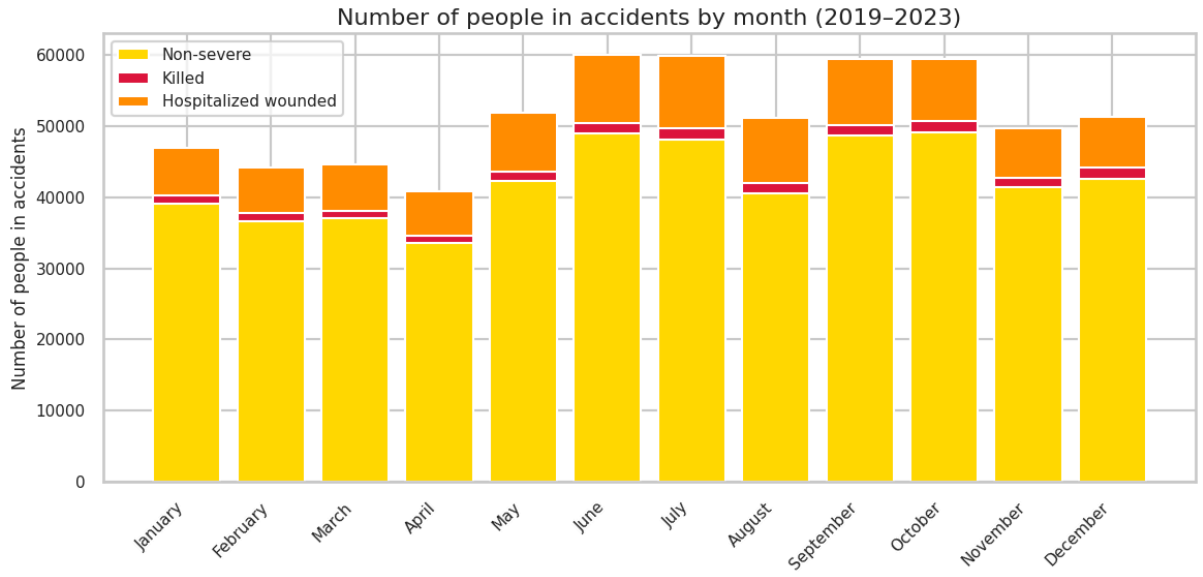
- date: combines day, month, and year into a standard date format.
- datetime: extends date by also including the hrmn field (hour and minute).
- hour: extracts the hour of the day from datetime, allowing us to analyze accident frequency by time of day.
- weekday: derives the day of the week from date, stored as a categorical variable for efficient grouping and clear visualization.

These transformations allow us to explore patterns such as whether accidents are more frequent on weekends, during specific weekdays, or at certain hours of the day.

3. Visualizations and statistics

3.1 Temporal analysis

3.1.1 By month



Description of the graph and analysis

The bar chart shows the number of people involved in accidents in France for each month (2019–2023), broken down into those in non-severe accidents, hospitalized, or killed. To test whether the likelihood of a severe accident (hospitalization or death) depends on the month, a binary logistic regression was performed with severe accident (yes/no) as the outcome, and January as the baseline. Cramér's V was also calculated to measure the overall strength of the association.

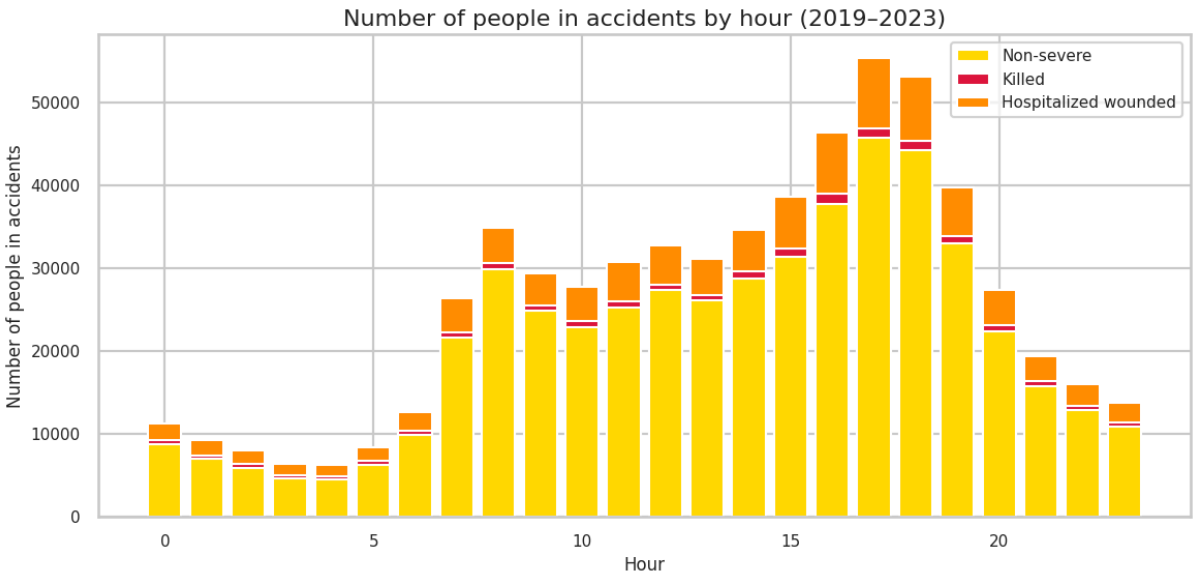
Results

Compared to January, the odds of a severe accident are about 22% higher in July and 31% higher in August.

Other spring and summer months (April, May, June, September, October) also show small but statistically significant increases.

The overall association between month and severity is weak (Cramér's V = 0.018), meaning that while the seasonal peaks are clear, most of the variation in accident severity is explained by other factors rather than the calendar month.

3.1.2 By hour



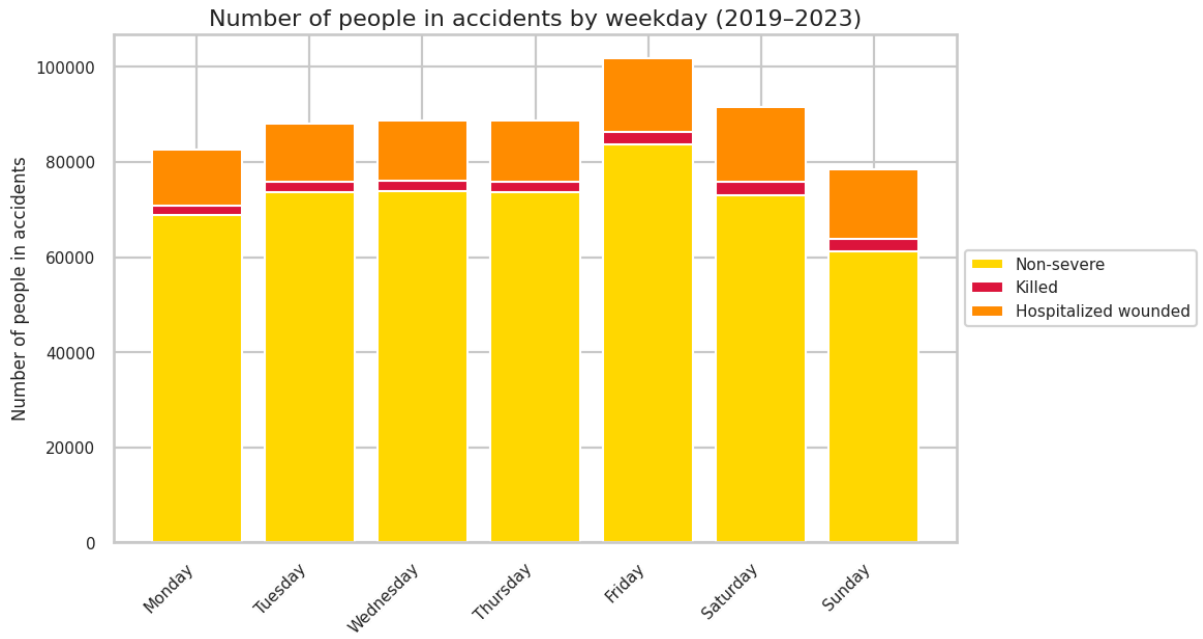
Description of the graph and analysis

Similar to the previous graph, the graph above shows the number of people involved in accidents in France for each hour of the day (2019–2023). It distinguishes between those in non-severe accidents, those hospitalized wounded, and those killed. Again, a logistic regression was performed with 0:00–0:59 as the baseline hour to test whether the likelihood of a severe accident (hospitalization or death) varies by time of day. Cramér's V was also calculated to measure the overall strength of the relationship.

Results

Severe accidents are most likely in the early morning hours (1:00-4:00 AM, up to 33% higher odds than midnight), while daytime and evening accidents are generally less severe. Time of day influences accident severity more than month does. Cramér's V = 0.047, is larger here than in the monthly analysis, but the effect is still weak. The results suggest that accidents at night and early morning, when conditions may be riskier (i.e, fatigue, alcohol, low visibility), are more likely to result in severe outcomes.

3.1.3 By weekday



Description of the graph and analysis

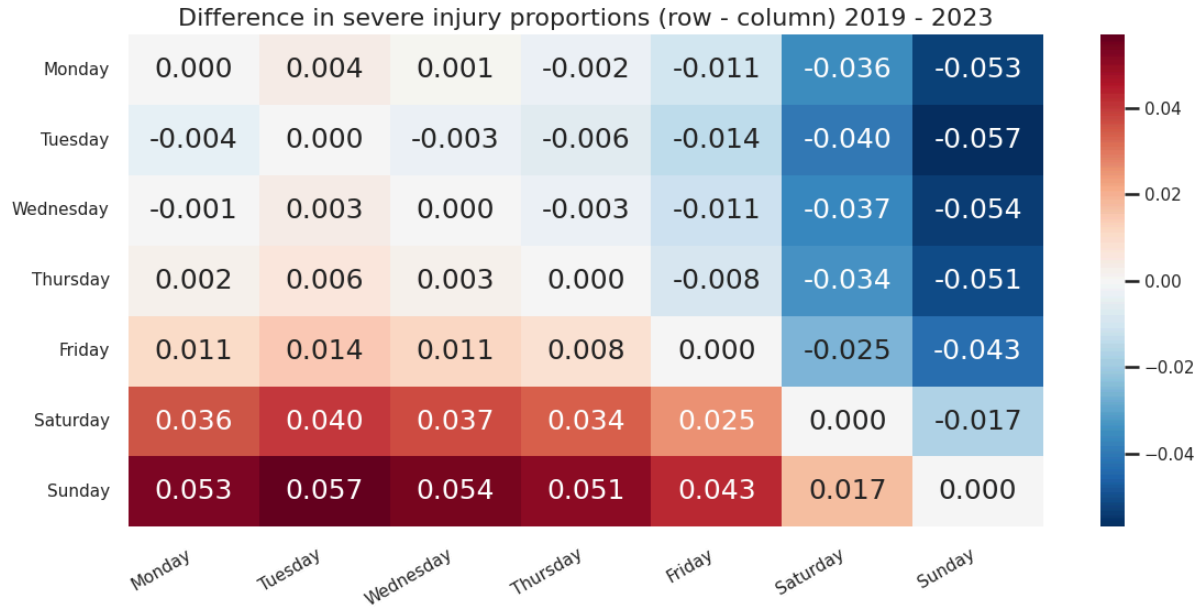
The analysis looked at how the day of the week relates to the severity of road accidents between 2019 and 2023.

The stacked bar chart shows how many people were involved in accidents each day, separated into three groups: Non-severe injuries, hospitalized wounded, and killed. Using a logistic regression model, we compared the odds of having a severe accident on each day to Monday (the baseline).

Results

Fridays had the highest total number of people involved in accidents, while Sundays had the fewest. However, the share of severe cases (hospitalized wounded or killed) was higher on weekends than on weekdays.

While weekday overall explains little variance in accident severity (Pseudo $R^2 \approx 0.0027$), the regression shows that weekends are riskier. On Saturday, the chance of a severe accident was about 27% higher than on Monday. On Sunday, the chance was about 41% higher than on Monday. Fridays also showed a smaller but still noticeable increase.



Description of the heatmap and analysis

The heatmap shows the differences in the proportion of severe injuries between each pair of weekdays. Each cell compares the day in the row with the day in the column:

Positive values (red shades) mean that the day in the row has a higher share of severe injuries than the day in the column.

Negative values (blue shades) mean the opposite – the row day has fewer severe injuries than the column day.

Values close to zero indicate little or no difference.

Results

The weekends (Saturday and Sunday) stand out clearly in dark red, showing they have higher proportions of severe injuries than any weekday.

Sunday has the highest severe injury rate overall, followed by Saturday.

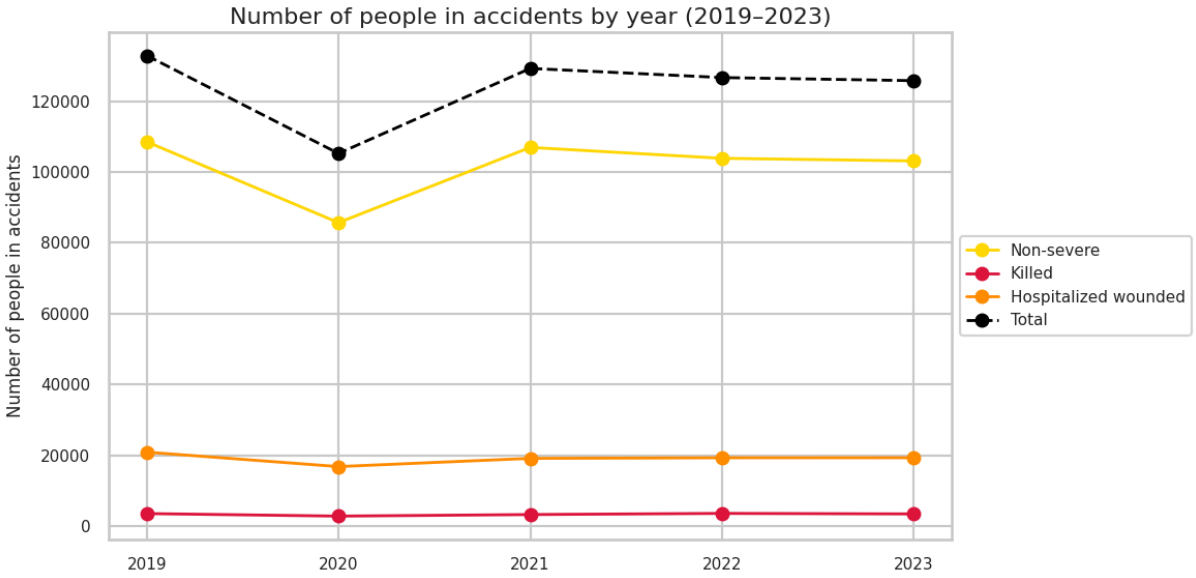
Monday to Thursday show very small differences between each other, suggesting injury severity is fairly similar across these days.

Friday shows a small increase compared to earlier weekdays, but still less severe than the weekend.

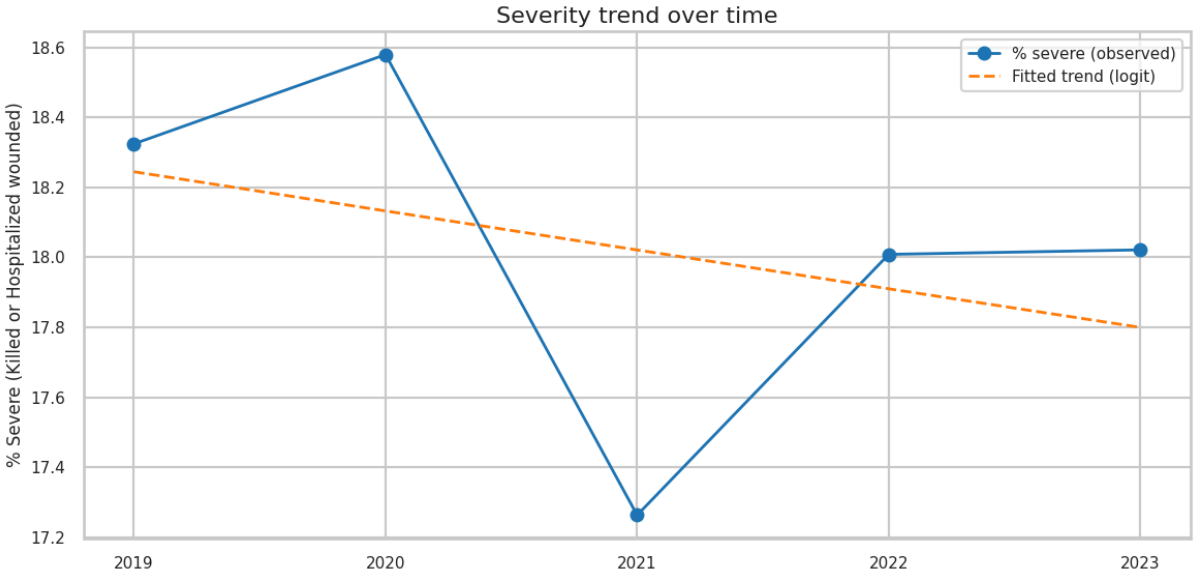
The largest differences occur when comparing Sunday vs. Monday–Thursday, where the proportion of severe injuries is around 5% higher on Sunday.

While accidents happen throughout the week, weekends stand out as riskier, with a higher likelihood of severe outcomes. This may be linked to factors like longer leisure trips, higher speeds, or alcohol use on those days.

3.1.4 By year



Year	Non-severe	Killed	Hospitalized wounded	Total
2019	108,527	3,497	20,852	132,876
2020	85,680	2,780	16,772	105,232
2021	106,936	3,219	19,093	129,248
2022	103,852	3,550	19,260	126,662
2023	103,120	3,398	19,271	125,789
Total	508,115	16,444	95,248	619,807



Description of the graphs and analysis

The first line chart shows the number of people involved in accidents each year (2019–2023), separated into three categories: Non-severe injuries, hospitalized wounded, and killed, along with the total number of people affected. The table displays the numbers for these categories as well. Overall, the number of people involved in accidents has decreased. The second chart tracks the percentage of people with severe injuries (those resulting in hospitalization or death) over time. A trend line from a logistic regression model estimates whether severity has increased or decreased across years. The analysis also calculated the odds ratio per year, which shows how the likelihood of a severe injury changes from one year to the next.

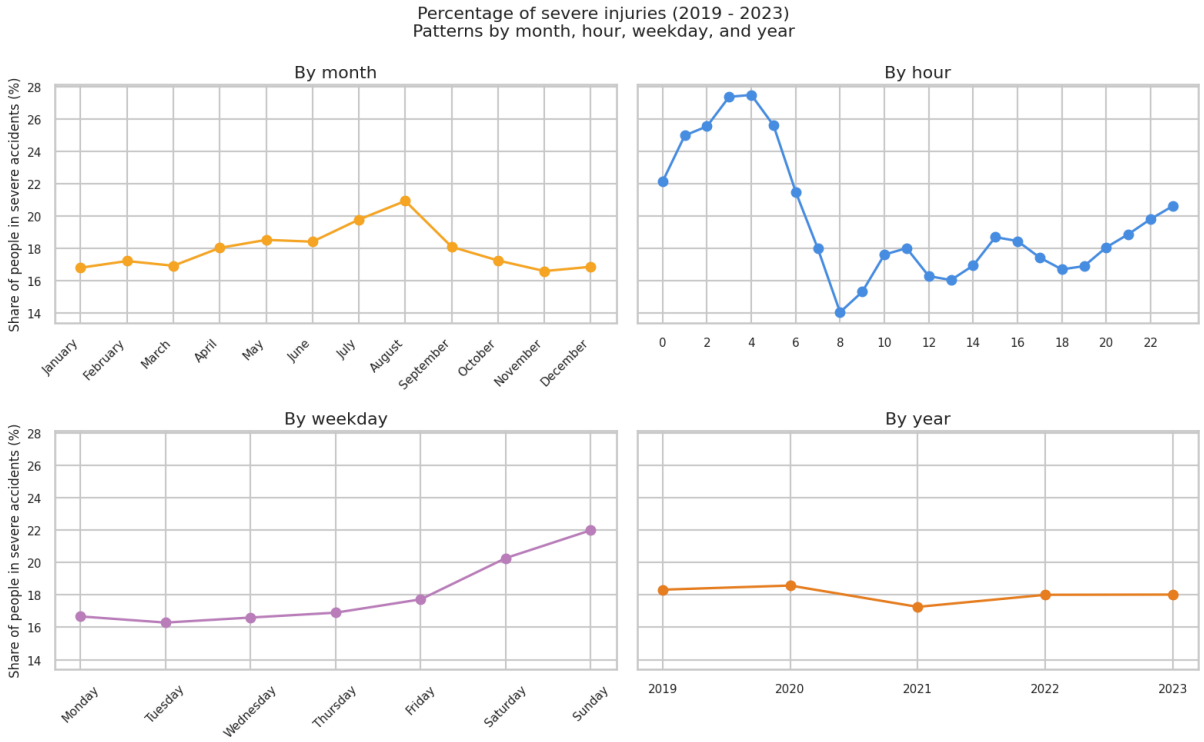
Results

Between 2019 and 2023, the proportion of people with severe injuries fell from 18.32% to 18.02%, a decline of 0.3 percentage points. The trend line confirms a gradual decrease in severity proportions over time.

The logistic regression found a per-year odds ratio of 0.992, meaning the odds of a severe injury decreased by about 0.8% per year. Over the entire 5-year period, that adds up to roughly a 4% total reduction in the odds of severe injuries.

The number of people in all three groups – non-severe, hospitalized wounded and killed – has decreased from 2019 to 2023. 2020 had fewer people in accidents overall, but a slightly higher proportion of people with severe injuries (18.58%) compared to the other years.

3.1.5 Month, hour, weekday and year



Description of the graphs and analysis

The figure shows the percentage of severe injuries (hospitalized wounded or killed) from 2019–2023, broken down by month, hour, weekday, and year.

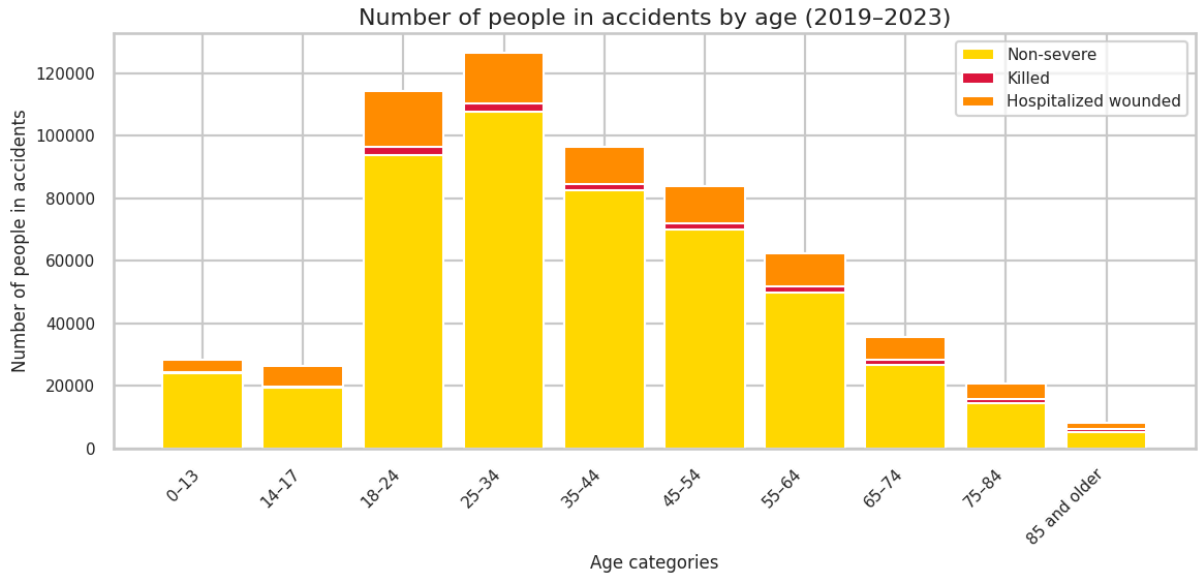
By month: the percentage of severe injuries is fairly stable (16–18%) for most of the year, with a peak in August (20.94%).

By hour: a strong pattern is visible. Severity is highest in the early morning (1:00–4:00 AM, up to 27.49%) and lowest in the morning at 8:00 AM (14.07%). It then gradually rises again in the evening.

By weekday: severity is lowest on weekdays (about 16%) but climbs on weekends, peaking on Sunday (22.00%).

By year: a small downward trend is visible, with severity proportions slightly decreasing from 18.32% in 2019 to 18.02% in 2023.

3.2 Age analysis



Description of the graph and analysis

The bar chart shows the number of people in accidents (2019–2023) by age group, broken down into non-severe, hospitalized wounded, and killed outcomes.

The largest groups involved in accidents are 18–24 and 25–34, followed by 35–44 and 45–54. Younger groups (0–13, 14–17) and the oldest groups (75–84, 85 and older) have fewer accidents overall. However, the proportion of severe accidents increases with age – especially in the oldest groups.

Results

We tested whether age and severity were related (Cramér's V) and then used logistic regression to measure how much more or less likely each age group is to be in a severe accident compared to children (0–13).

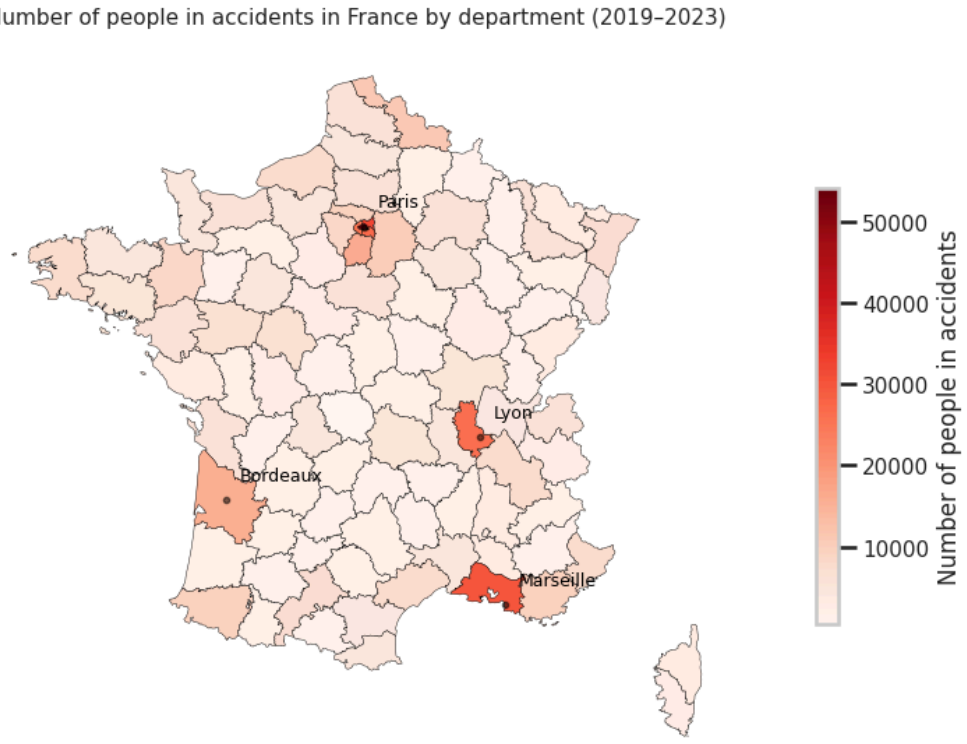
Cramér's V at 0.094 shows a stronger association than month, weekday or hour effects.

Accident counts are highest among young adults (18–34), but severity is not highest there. Severity risk rises with age: while older adults have fewer accidents, those accidents are much more likely to be severe.

Teenagers (14–17) also stand out with higher severity risk, possibly due to inexperience. The strongest severity risk group is 85 and older, with more than three times the odds of severe accidents compared to children.

3.3 Location analysis

3.3.1 Regional accident distribution



Note: Map includes metropolitan France only. Overseas departments (971–988) excluded from visualization.

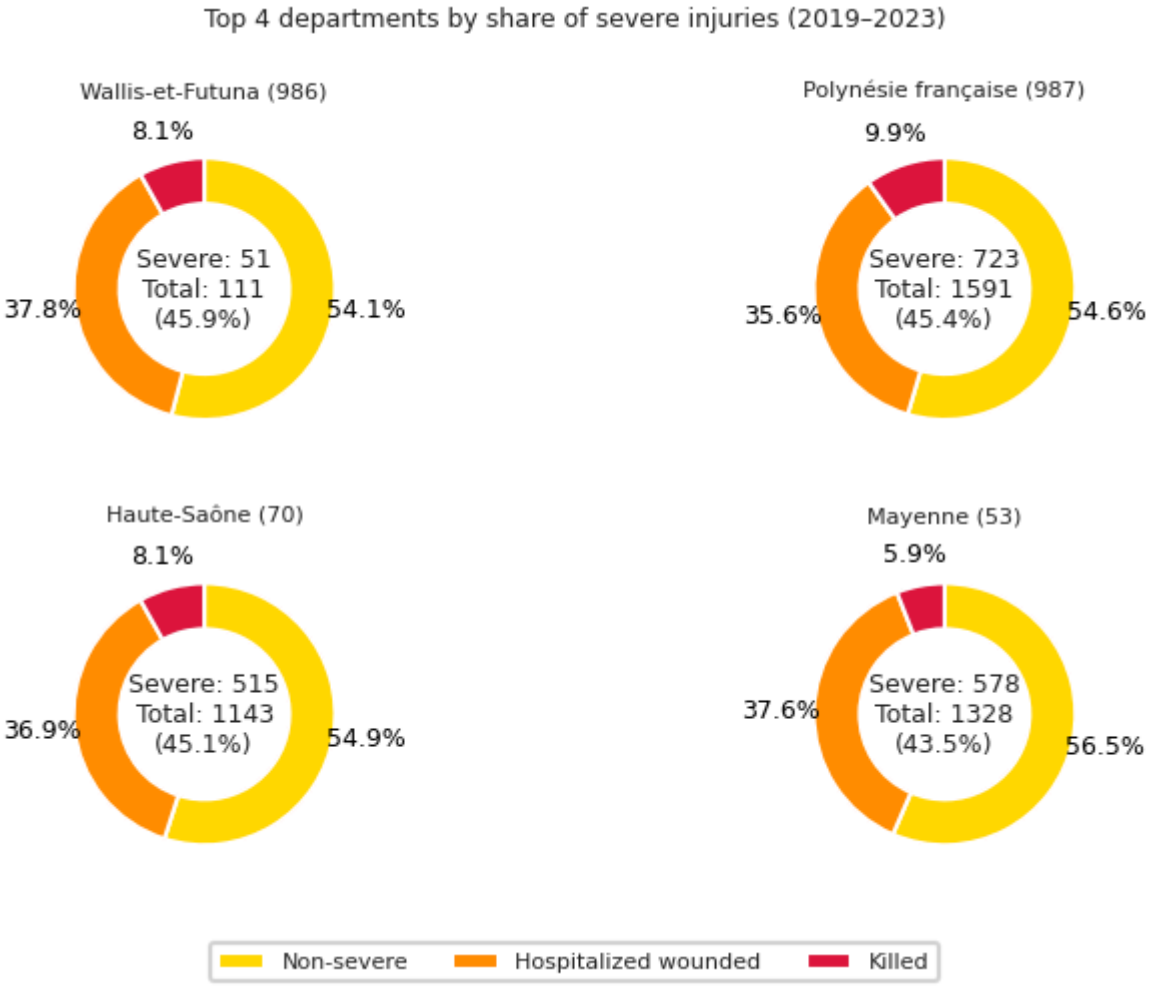
Description of the graph and analysis

The map displays the total number of people involved in road accidents across French departments (2019–2023). Each department is colored according to the total number of accident participants, with darker shades of red indicating higher counts.

The data combines all severity levels (non-severe, hospitalized wounded, and killed) into a single total per department. Departments around Île-de-France (Paris), Rhône (Lyon), Bouches-du-Rhône (Marseille), and Gironde (Bordeaux) show the highest number of people in accidents.

In contrast, rural or less densely populated regions, particularly in central and western France, display fewer people involved in accidents.

3.3.2 Departments with the highest share of severe injuries



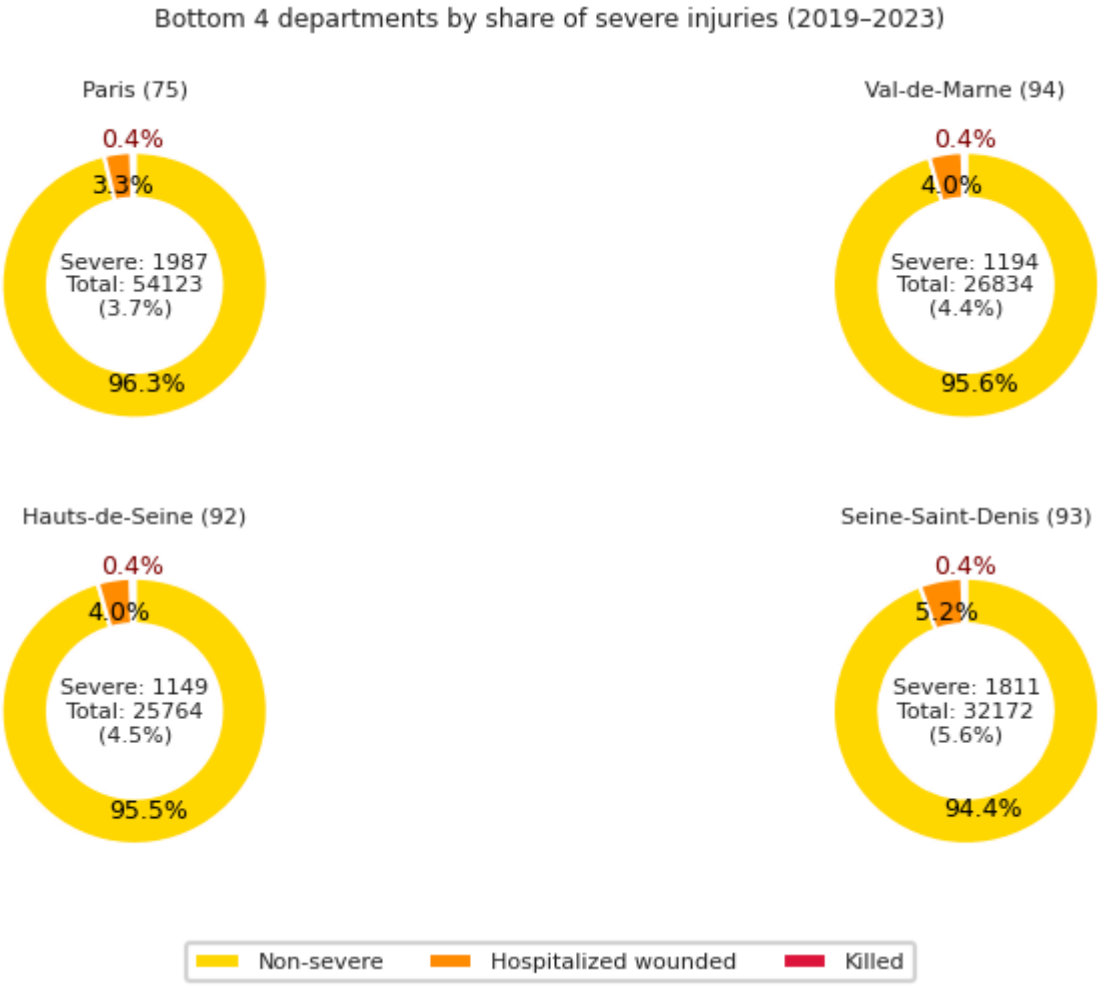
Description of the graphs and analysis

The chart presents the four departments with the highest proportion of severe road injuries (Hospitalized wounded and Killed combined) between 2019 and 2023. Each donut chart shows the share of Non-severe (yellow), Hospitalized wounded (orange), and fatal (red) injuries.

Departments from French overseas territories — Polynésie française (987) and Wallis-et-Futuna (986) — appear among the top four, alongside Haute-Saône (70) and Mayenne (53) in mainland France.

In all four regions, severe cases represent over 43% of all recorded accidents, with hospitalized injuries forming the majority of these severe outcomes.

3.3.3 Departments with the lowest share of severe injuries



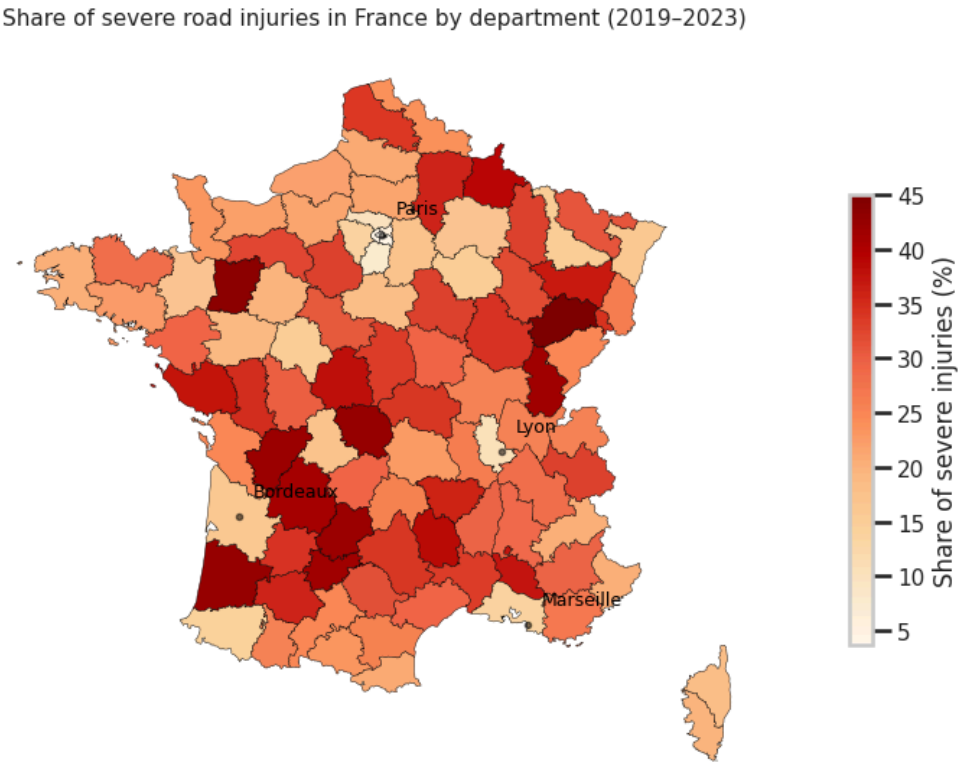
Description of the graphs and analysis

This figure displays the four departments with the lowest proportion of severe road injuries (hospitalized and fatal cases combined) between 2019 and 2023. Each donut chart shows the distribution of non-severe (yellow), hospitalized wounded (orange), and fatal (red) accidents, alongside total accident counts and the share of severe outcomes.

The departments shown — Paris (75), Val-de-Marne (94), Hauts-de-Seine (92), and Seine-Saint-Denis (93) — are all part of the Île-de-France region, highlighting a strong geographical concentration in the capital area.

Across these four departments, severe injuries account for only around 4–6% of all recorded accidents, which is substantially lower than the national average (18.02%).

3.3.4 Regional distribution of severe road injuries



Note: Map includes metropolitan France only. Overseas departments (971–988) excluded from visualization.

Description of the graph and analysis

The map shows the percentage of severe road injuries (hospitalizations and deaths) among all accidents across French departments from 2019 to 2023. A Chi-squared test was used to examine whether these shares differ significantly between regions, and Cramér's V measured the strength of this association.

Darker shades indicate departments with a higher share of severe injuries, while lighter shades represent lower shares. Higher values appear mainly in rural areas, whereas major urban centers such as Paris, Lyon, Marseille, and Bordeaux show lower proportions.

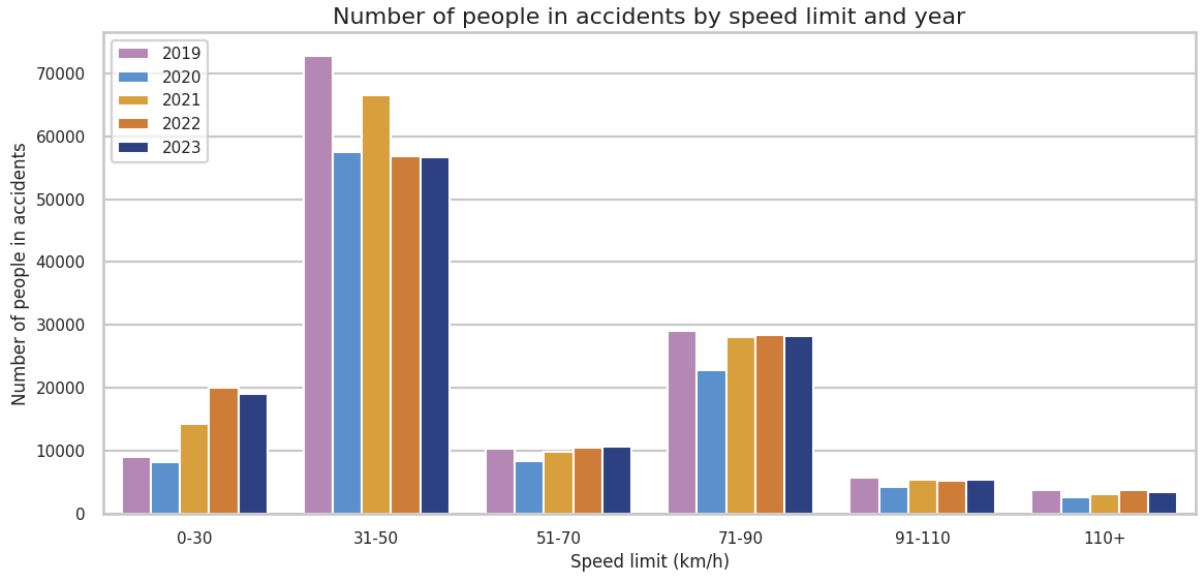
Results

The Chi-squared test revealed significant regional variation ($p < 0.001$). Cramér's V = 0.27 indicates a moderate association, meaning that accident severity is not evenly distributed across departments.

Clear geographical patterns emerge: Departments in rural areas – particularly in the northeast and southwest – show higher proportions of severe injuries, while large metropolitan regions (Île-de-France, Lyon, Marseille, Bordeaux) display lower shares. This suggests that accidents in rural regions are more likely to lead to hospitalization or death, possibly due to higher driving speeds/speed limits, longer emergency response times, and differences in infrastructure.

3.4 Speed limit analysis

3.4.1 Number of people in accidents by speed limit and year



Description of the graph and analysis

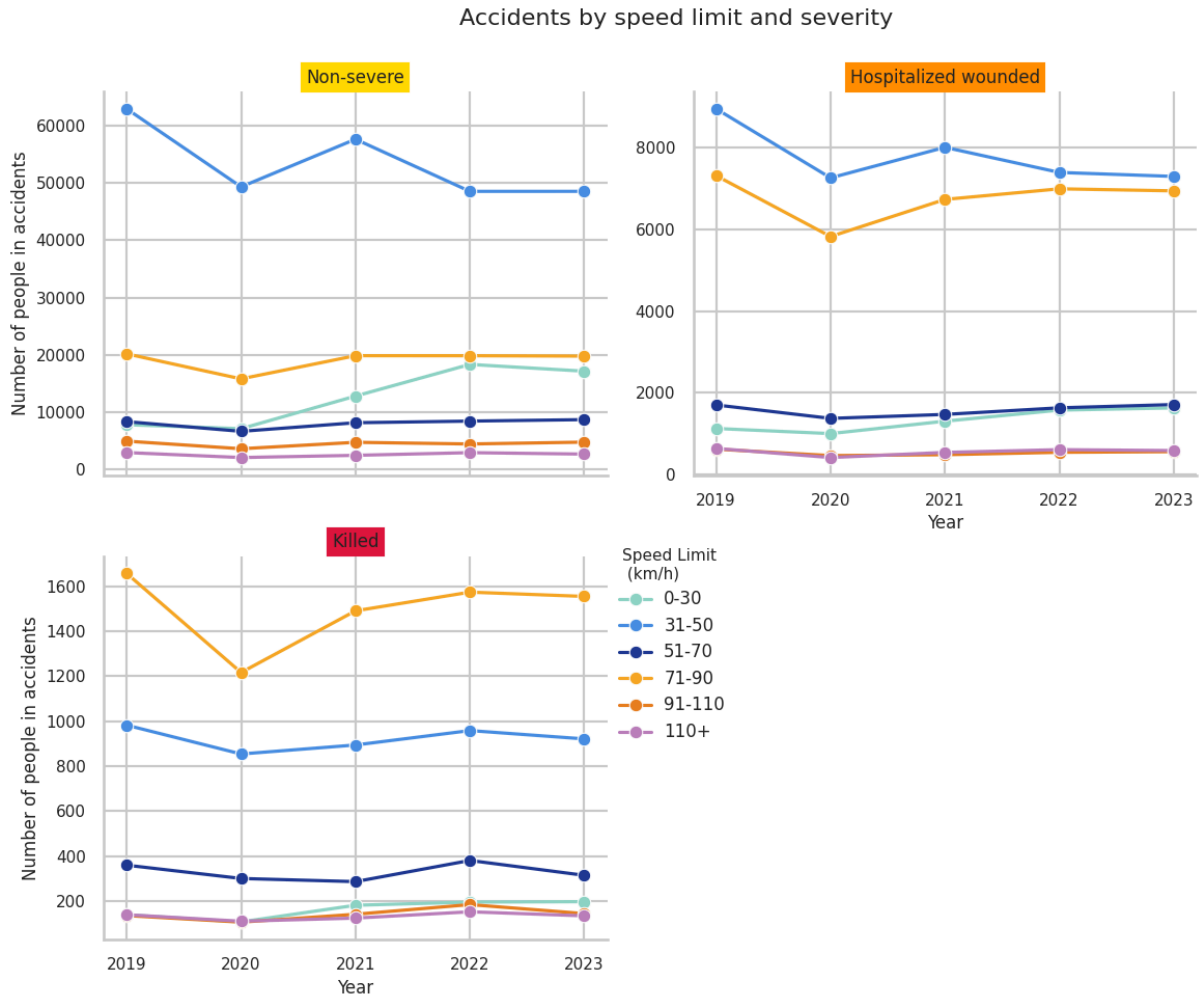
This bar graph represents the total number of people involved in accidents between 2019 and 2023, grouped by the speed limit range (0–30, 31–50, 51–70, 71–90, 91–110, 110+ km/h). To differentiate the years we chose different colors, allowing for comparison over time.

Results

From this graph we notice that accidents are most frequent on roads with speed limits ranging from 31-50 km/h and 71-90 km/h, which correspond mainly to urban and peri-urban areas. We see the number of accidents significantly decrease at higher (above 90 km/h) and medium speed limits (51-70 km/h). Across years, a clear dip is visible in 2020 due to reduced mobility during COVID-19 restrictions, followed by an increase from 2021 to 2023.

Overall most accidents occur in areas with moderate speed limits, confirming that denser traffic conditions and mixed road usage contribute more to accident frequency than high-speed environments. However these results do not take into account the injury severity.

3.4.2 Number of people in accidents by speed limit and injury severity



Description of the graphs and analysis

These three graphs represent the number of people involved in accidents across speed limits though this time taking into account injury severity. Each subplot also shows the evolution from 2019 to 2023. To further analyze we tested whether accident severity is independent of speed limit by performing a Chi-square test. Cramer's V was calculated to gauge how strong the relationship is. Standardized residuals were used to identify which speed ranges and severities are most important in this relationship.

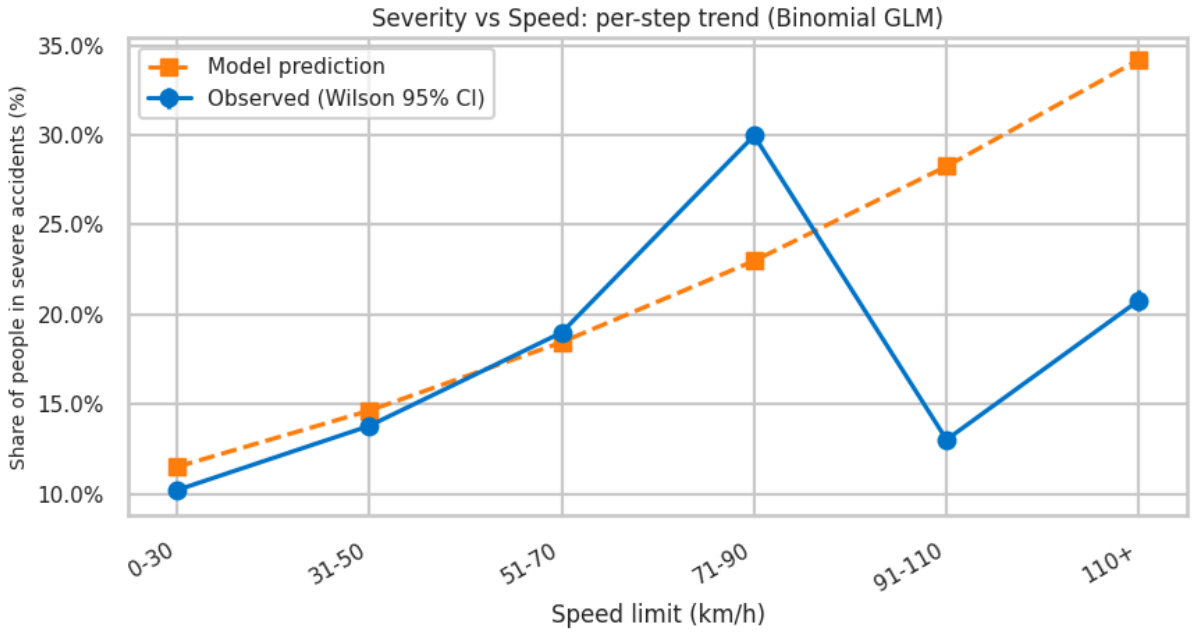
Results

There is a statistically significant relationship between the posted speed limit of a road and the severity of injuries in accidents.

The strength of this relationship (Cramer's V = 0.134) is weak, meaning the link is real but not solely determined by speed – other factors (like road type, traffic control, and vehicle mix) likely contribute. Nevertheless, the analysis clearly shows that speed limits between 31-70 km/h are associated with a higher proportion of non-severe and hospitalized wounded cases, likely due to their higher traffic exposure (most likely urban and peri-urban environments). However, high-speed roads (91+ km/h) have less total accidents though showed far more severe and

fatal injuries than expected, with particularly high residual values for “Hospitalized wounded” and “Killed.”
This indicates a clear pattern: as the posted speed limit increases, the likelihood of serious or fatal outcomes rises.
In contrast, lower-speed roads tend to have less severe accident outcomes.

3.4.3 Severity vs. speed: per step trend

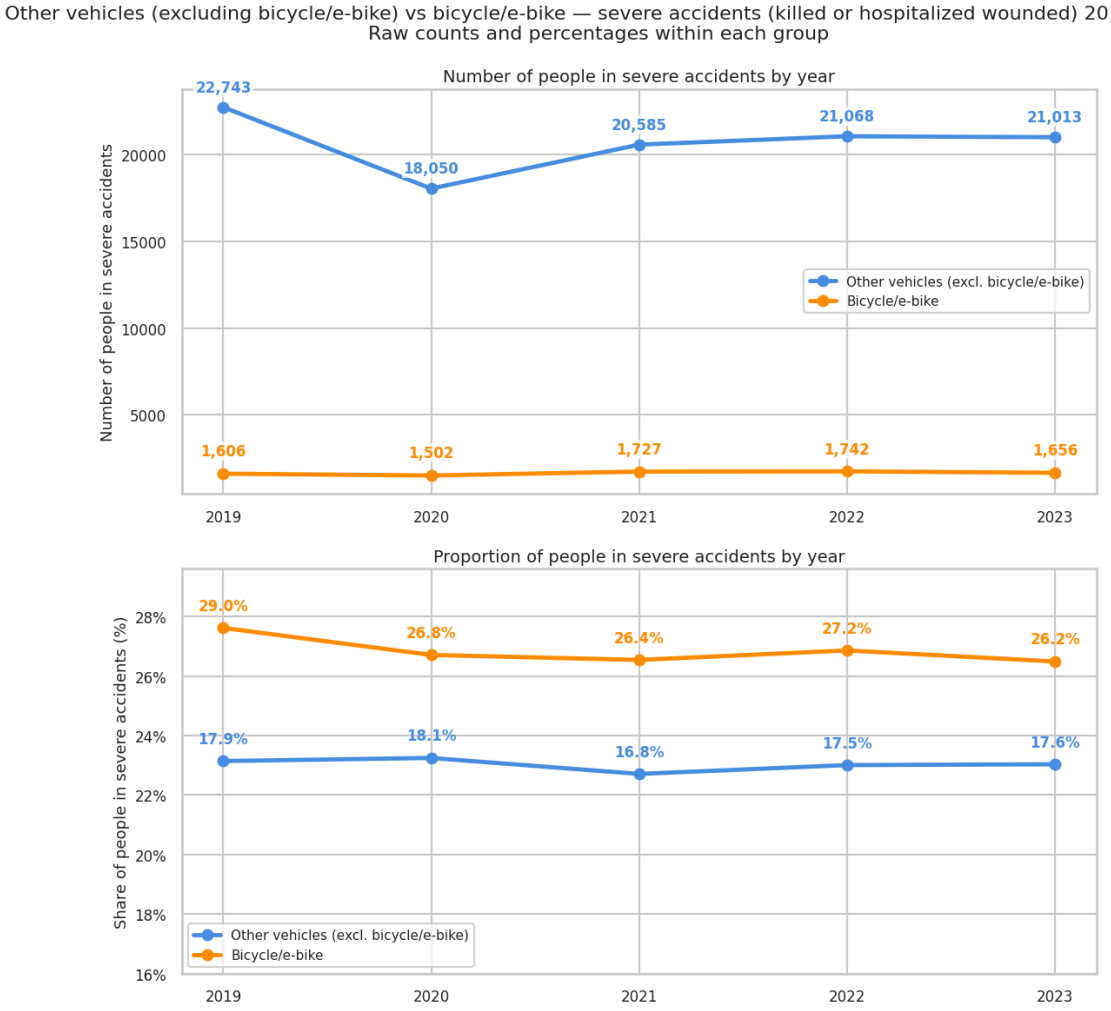


Description of the graph and analysis
This line chart shows the percentage of severe accidents (hospitalized wounded or killed) for each speed limit range. The solid line represents observed data, while the dashed line corresponds to a Binomial Generalized Linear Model (GLM) prediction with 95% confidence intervals. This model helps us estimate the likelihood of a severe accident depending on the speed limit.

Results
We see that the probability of a severe accident increases steadily with speed limit. There is a big spike in the 71–90 km/h range, where around 30% of accidents result in hospitalization or death. At lower speed limits (0–30 km/h), severity remains small, probably reflecting less impact force. Interestingly for higher speed limits beyond 90 km/h, the severity ratios (proportion of accidents that result in hospitalized wounded or killed) compared to the total number of accidents decline, possibly due to improved road infrastructure and controlled traffic on highways. This trend indicates that roads with speed limits around 71–90 km/h represent the highest relative risk, combining both frequent use and severe outcomes.

3.5 A Special Case: Bicycles/e-bikes and Severity

3.5.1 Comparison of bicycle and e-bike users with other road users



Description of the graphs and analysis

The charts compare all other vehicle accidents with those involving bicycles and e-bikes from 2019 to 2023, focusing on severe outcomes (killed or hospitalized wounded).

The top chart shows the number of people in severe accidents per year for all vehicles (excluding bicycles and e-bikes) and for the bicycle and e-bike users.

The bottom chart shows the percentage of people in severe accidents for the same groups, highlighting how severe outcomes differ between the groups over time.

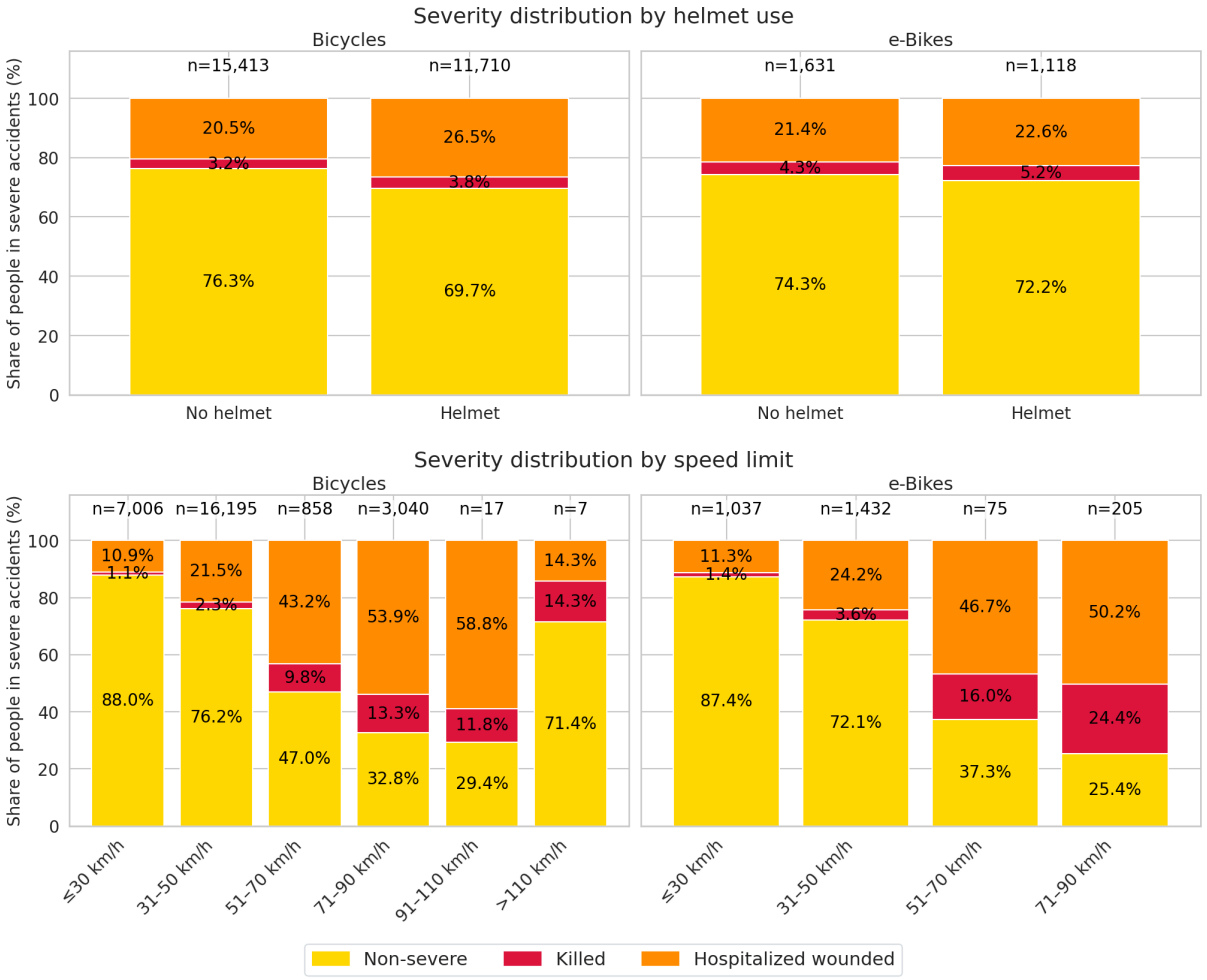
To statistically test whether bicycle and e-bike users are more (or less) likely to have a severe outcome than other road users, and how trends have changed, a logistic regression model was run.

Results

From 2019 to 2023, bicycle and e-bike accidents had a consistently higher proportion of severe injuries than those involving other vehicles. In 2019, 29.0% of bicycle and e-bike injuries were severe compared with 17.9% for all other vehicles, and the odds of a severe outcome were 81% higher for bicycles and e-bikes compared to other vehicles. Across all years, cyclists consistently have a higher predicted probability of a severe accident than other vehicle users.

While the overall share of severe injuries declined for both groups, the difference remained clear: by 2023, 26.2% of bicycle and e-bike injuries were severe versus 17.6% for all other vehicles. Absolute numbers are much lower for bicycle and e-bike users compared to all vehicles, however the proportion of severe outcomes is consistently higher (around 26–29% vs. 17–18%). This highlights that cyclists and e-bike users face a disproportionately high risk of severe injury or death when involved in accidents as compared to other road users.

3.5.2 Bicycle and e-bike users: helmets and speed limits



Description of the graphs and analysis

The graphs show how the severity of bicycle and e-bike accidents varies by helmet use and by speed limit on the road where the accident occurred. Each stacked bar represents the percentage of people within that group resulting in: Non-severe injuries, hospitalizations, and fatalities.

What we tested:

a. Helmet vs no-helmet within each group (bicycles and e-bikes). We compared the share of accidents that were severe (killed or hospitalized). Instead of only asking "is it significant?" we reported how big the difference is (percentage-point gap, risk ratio, odds ratio) and a confidence interval.
b. Speed environment within each group: we checked whether severity rises as the posted speed gets higher. We summarized the change per step up the speed scale (i.e. from 31–50 to 51–70 km/h).

Because the dataset is very large, significance tests will often be tiny; the focus here is on effect size and direction.

Results

For bicycles, riders who wore helmets had a higher share of severe injuries (30.3%) than those without helmets (23.7%). This difference is small but statistically clear. It may reflect that helmet users are more often involved in faster or riskier situations rather than helmets increasing risk.

For e-bikes, helmeted riders also had a slightly higher share of severe cases (27.8% vs. 25.7%), but the difference was very small and only weakly significant.

Speed showed the strongest effect on injury severity for both bicycles and e-bikes. Each step up in the speed limit category (for example, moving from 30–50 km/h to 50–70 km/h) more than doubled the odds of an accident leading to hospitalization or death.

Speed environment is a strong driver of severity (large, monotone effect), and the deviations hint at road-type differences worth exploring.

Part B: advanced dataviz and dashboarding report

1. Classification of the problem

1.1 Context

This part of our project aims to primarily permit users to understand how different criterias/factors have an impact on the number of people in road accidents or severity of road accidents in France in an intuitive and interactive way. This dataset contains high-quantity, complex information (i.e. age, road category, speed limit, user type) which is difficult to interpret from just graphs alone.
With the use of visualization tools we want to address how to transform a large, multidimensional dataset and advanced statistics into a clear, understandable, and visual story. We want to achieve this by highlighting relationships between factors, identifying the areas and driving conditions with higher severity risks, and pointing out trends.

1.2 Objectives

Our visualization goals are the following:

- Design an interactive Power BI dashboard that enables users to explore road accidents in France and measure number/severity by the factors available to us.
- Present key findings through dynamic visuals, color-coded severities (i.e. color yellow for non-severe injuries), and clear hierarchies that guide users from main elements to detailed ones.
- Incorporate interactivity and storytelling elements (filters, slicers, and tooltips) to make the dashboard more engaging and exploratory.
- Ensure consistency in our chart types, and accessibility by providing a readable format, and meaningful legends that maintain clarity even with large datasets.

To reach our objectives we set the following performance metrics/evaluation criteria to assess our effectiveness:

- Consistency and visual coherence: We assessed color palettes, font choices (and size), and layout alignment to ensure all pages followed a unified visual identity.
- Data accuracy and validation: We compared our Power BI visuals and tables against results obtained in Python (Colab) to confirm that all figures and calculations matched across platforms.

- Interactivity and usability: We jointly tested the behavior of slicers, filters, and tooltips to ensure they worked logically and helped users explore the data smoothly.
- Structural organization: We created schema outlines to plan each page's content and theme, ensuring that the overall dashboard flow made sense and that each page focused on a specific analytical angle (e.g., demographics, severity, speed, or location) to better best represent the story we wanted to tell about road accidents in France.

2. Dashboard design and optimization

2.1 Data model and relationships

Before building the Power BI dashboard, we designed a relational data model to ensure consistent structure, efficient aggregation, and clear analytical logic. The model follows a star schema, with a central Fact table surrounded by several Dimension tables. This approach enables fast filtering and cross-analysis while avoiding redundancy. All model tables (except the Date Dimension) were prepared in Python and exported with only the required columns, correct data types, and pre-binned/grouped fields to minimize the Power BI file size and improve refresh and query performance. The date dimension was created directly in Power Query to guarantee a single, consistent calendar and robust joins with the fact table. Building it upstream ensures stable keys, correct data types, and avoids ambiguity from time components, which can otherwise lead to broken or inactive relationships in the model.

The Fact table (Fact_User) contains one row per accident participant and combines information from the four raw sources: Characteristics, Places, Vehicles, and Users. The included columns are either a foreign key (FK) linked to the corresponding dimension table or categorical variables relevant for analysis:

Foreign keys

- num_accident - Accident identifier
- date_key - Dim_Date[date_key]
- age - Dim_Age[age]
- dep - Dim_Department[dep_code]
- speed_limit - Dim_SpeedBand[speed_limit]
- injury_severity - Dim_Severity[injury_severity]

Analytical attributes

- hour (0–23)
- category_user (driver, passenger, pedestrian)
- gender (male, female, not specified)
- has_helmet (helmet, no helmet)
- category_vehicle (bike, e-bike, other vehicle)
- road_category (Highway, National road, Departmental road, Communal way, Urban metropolitan road, Other)

The following Dimension tables were created to support flexible slicing and filtering in Power BI:

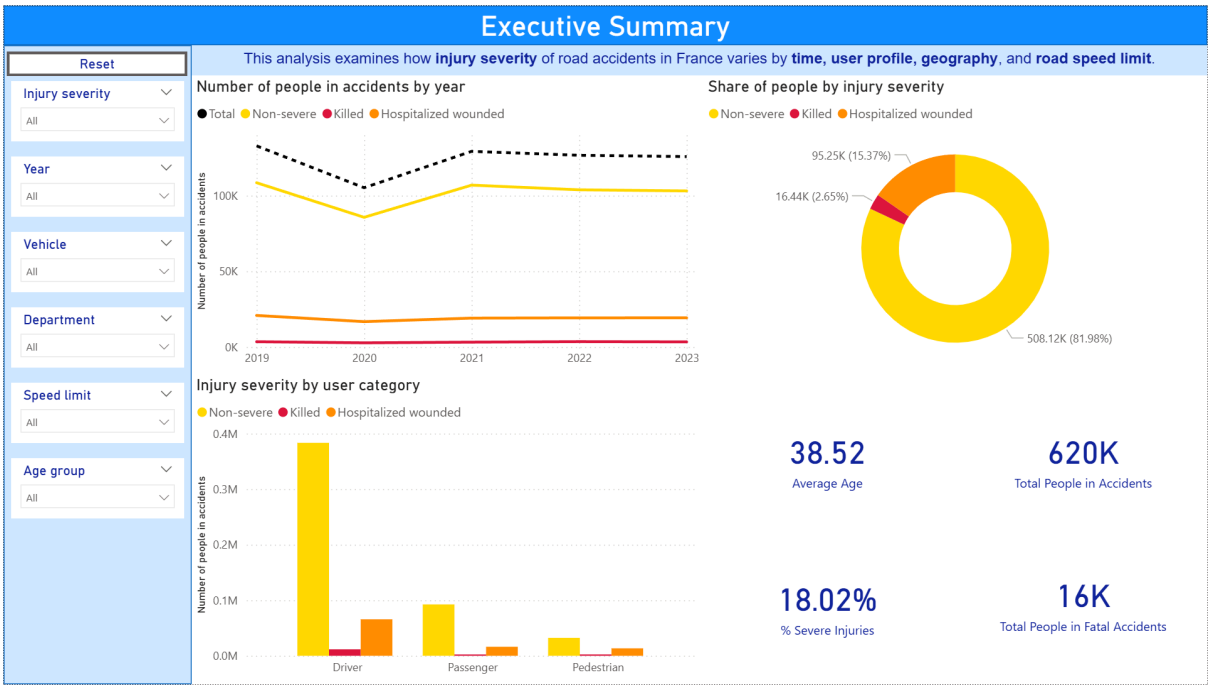
- Dim_Date – derived from the date_key (date of the accident); contains date_key (Primary Key = PK), year, month_number, month, month & year ID, month & year, day, weekday_number, weekday, quarter, quarter & year ID, and quarter & year to enable detailed temporal analysis (seasonality, trends, and time-based aggregations).
- Dim_Age – derived from age (users age); contains age (PK), age_bucket_label (i.e. “0–13”, “14–17”, ..., “85 and older”) and sort_order (1 - N to sort by age_bucket_label) to support demographic analysis
- Dim_Department – derived from the dep (department code); contains dep_code (PK; i.e. “75”, “2A”, “971”), dep_name, dep_code_name (Paris (75)) and dep_name_fr (used for better assignment for the map visual) to support geographic filtering and mapping.
- Dim_SpeedBand – derived from speed limit; contains speed_limit (PK; i.e. 30, 50, 100, 110), speed_band (i.e. “0–30”, “31–50”, “91–110”, “110+”), sort_order (1 - N to sort by speed_band) and model_prediction to compare severity across speed environments.
- Dim_Severity – derived from injury_severity; contains injury_severity (PK), injury_severity_name (Not Specified, Unscathed, Light Injury, Hospitalized Wounded, Killed), severity_name_order (1 - N to sort by in injury_severity_name), severity_label (“Non-severe”, “Hospitalized wounded”, “Killed”), severity_label_order (1 - N to sort by in severity_label), is_severe (“Non-severe”, “Severe”) and color (for explicit color coding) to standardize severity groupings and visual order.

All relationships are one-to-many, with the dimension table on the “one” side, ensuring proper filter propagation throughout the model.

This model structure offers several advantages:

- Clear separation between facts and descriptive attributes.
- Consistent filtering across all visuals (via one-to-many relationships).
- Optimized performance for large datasets (over 600 000 rows).
- Flexibility to add future dimensions such as Dim_Holiday (to study the impact of public holidays) without redesigning the schema.

2.2 Executive summary page



Description of the page

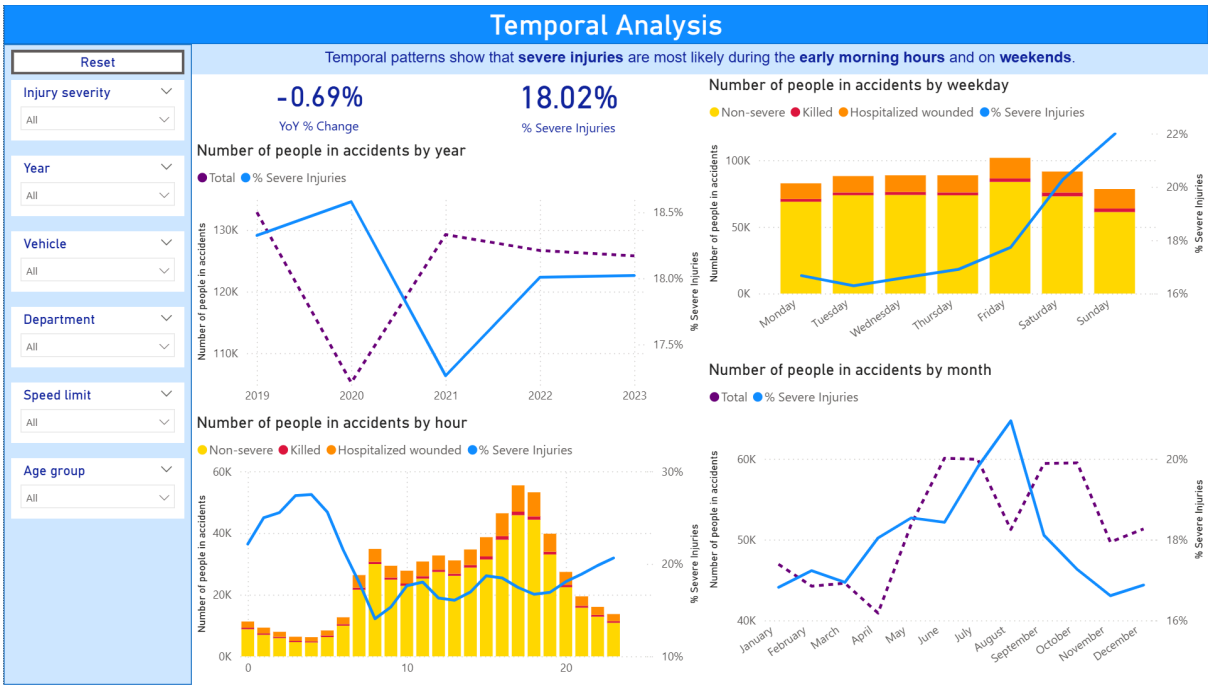
This overview page highlights the overall scale and severity of accidents across years. It summarizes key metrics such as total people involved, share of severe injuries, and fatal accidents. Line and donut charts visualize trends in injury severity and year-over-year developments, offering an at-a-glance understanding of national accident patterns.

Key takeaways

The line chart shows the number of people involved in accidents each year (2019–2023), separated into three categories: Non-severe injuries, hospitalized wounded, and killed, along with the total number of people affected. Overall, the number of people involved in accidents has decreased.

The ratio of non-severe to severe injuries is about four to one (about 80 % to 20%) and we see a similar ratio by user category. While most users in this dataset represent drivers, both drivers and passengers have about a four to one ratio of non-severe to severe-accidents. However, for pedestrians, the ratio of non-severe to severe is about three to one, meaning pedestrians comparatively face a higher risk of severe injury than drivers or passengers.

2.3 Temporal analysis page



Description of the page

This page examines time-based dynamics of accidents. It displays variations by year, month, weekday, and hour, revealing when accidents are most frequent and severe. Patterns such as weekend peaks or rush-hour clusters help identify high-risk time periods and support preventive planning.

Key takeaways

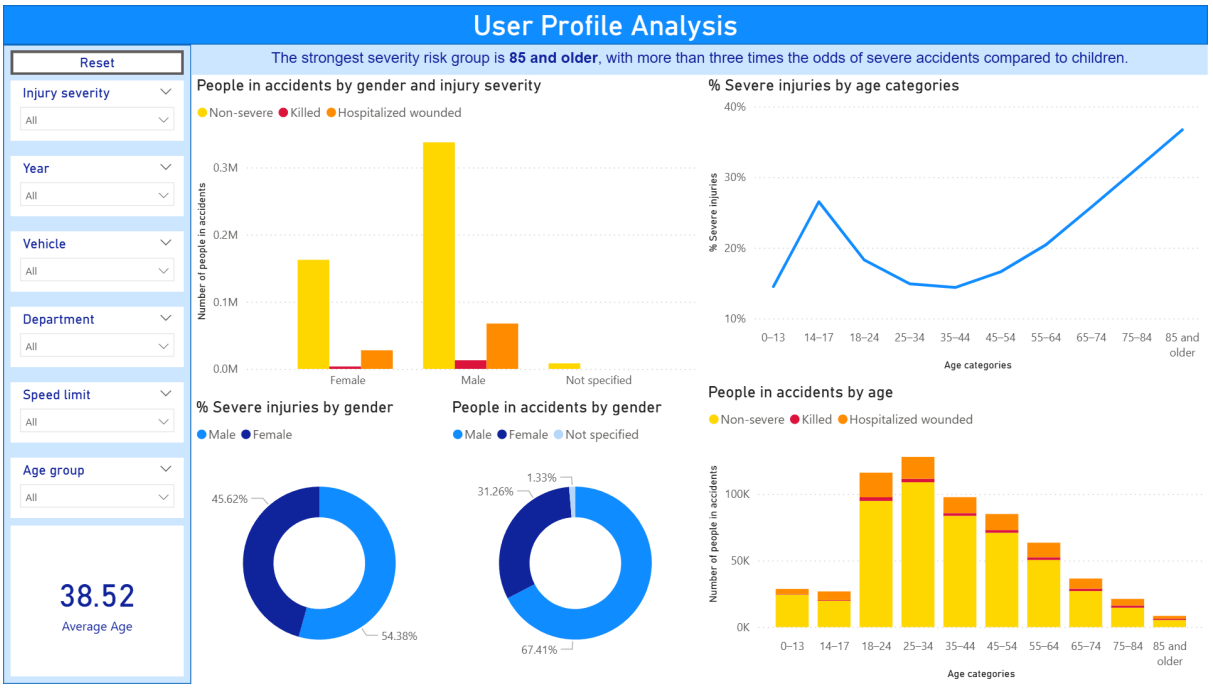
Over time, a small but consistent decline in injury severity was observed: from 18.32% in 2019 to 18.02% in 2023. The logistic regression found a per-year odds ratio of 0.992, meaning the odds of a severe injury decreased by about 0.8% per year. Over the entire 5-year period, that adds up to roughly a 4% total reduction in the odds of severe injuries, suggesting gradual improvements in road safety and medical response.

Across the week, weekends clearly stand out. Sundays have the highest share of severe injuries (22%), followed by Saturdays, while weekdays show little variation (around 16%). Regression analysis confirmed that the risk of severe injury or death is 27–41% higher on weekends than on Mondays, likely reflecting higher speeds, leisure travel, and alcohol consumption.

Time of day showed stronger patterns. Early morning hours (1:00–4:00 AM) have the highest severity rates – up to 33% higher than at midnight – likely linked to fatigue, alcohol use, and low visibility.

Injury severity varies modestly across months, peaking in July and August, when odds of a severe injury are 22–31% higher than in January. However, the overall association between month and severity is weak, suggesting that seasonal effects are real but limited.

2.4 User profile page



Description of the page

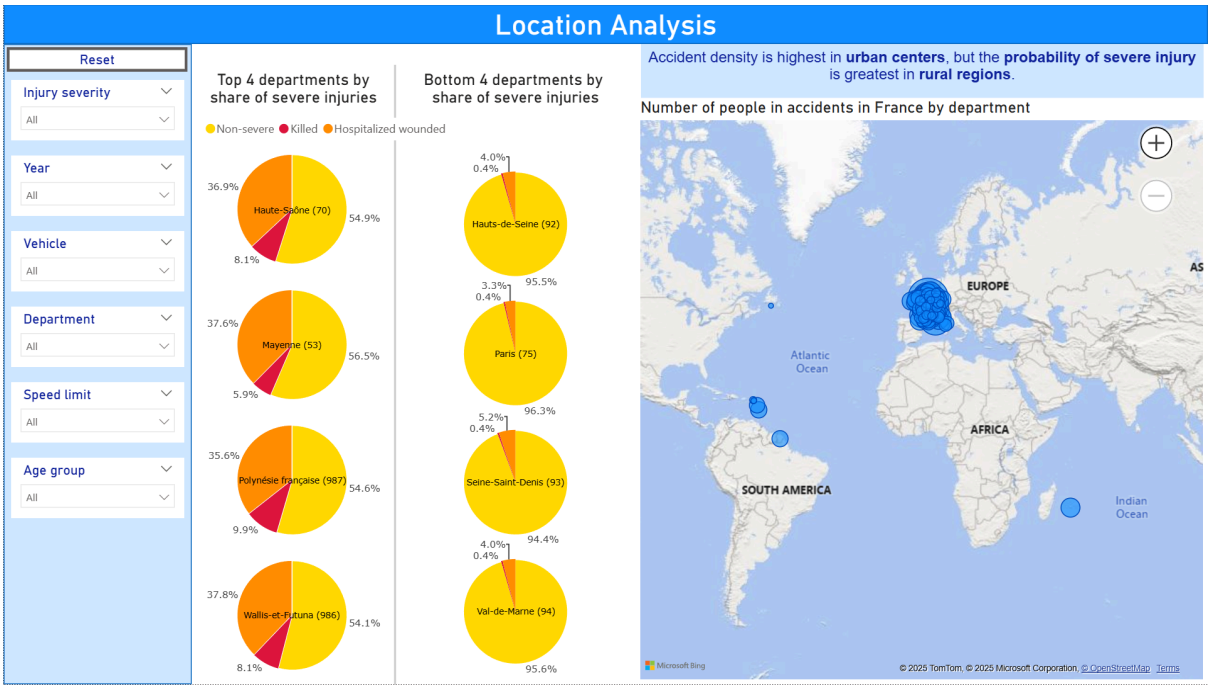
Focusing on demographics and user roles, this page compares accidents across gender and age group. It highlights severity differences – e.g., higher severity risks among elderly users – and visualizes how risk exposure changes across population groups.

Key takeaways

Severity increases with age. Although young adults (18–34) have the most injuries, older adults (especially 85 and older) face far higher odds of severe outcomes – over three times higher than for children. Teenagers (14–17) also show elevated severity, possibly reflecting inexperience.

When looking at gender, males represent roughly two-thirds of all individuals involved in road accidents between 2019 and 2023. They are not only more frequently involved in accidents but also display a slightly higher proportion (54% vs. 46%) of severe outcomes than females. Although the gap is not extreme, it reinforces the importance of considering gender when analyzing accident trends and severity patterns.

2.5 Location page



Description of the page

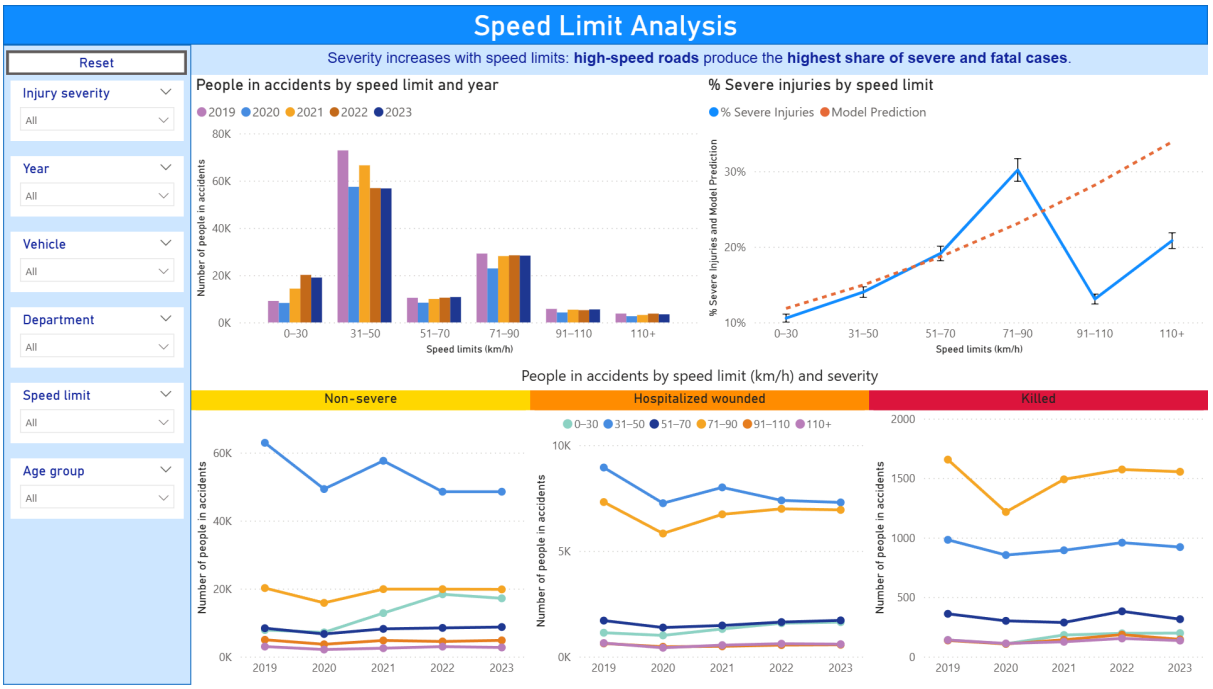
This page explores geographical differences across French departments. It ranks the top and bottom regions by share of severe injuries and maps the spatial distribution of accidents across France and overseas territories. Regional variations help pinpoint local safety challenges and contextual factors like urban density or infrastructure.

Key takeaways

Spatial patterns are striking. Urban and metropolitan regions (especially Île-de-France) record the most injuries, but their severity rates are the lowest (only 4–6% severe cases). In contrast, rural and overseas departments such as Polynésie française, Wallis-et-Futuna, Haute-Saône, and Mayenne show over 43% severe cases.

Statistical testing confirmed a moderate spatial association (Cramér's V = 0.27), meaning location influences severity – likely due to speed, emergency response times, and road infrastructure.

2.6 Speed limit page



Description of the page

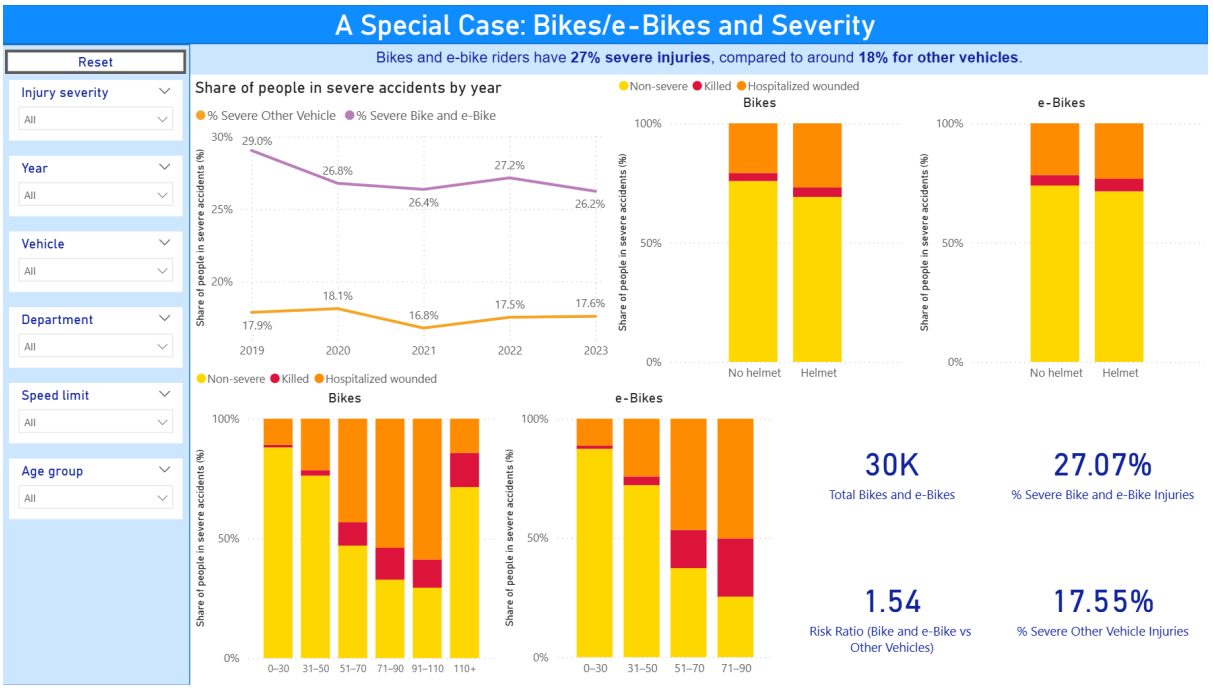
Here, the dashboard investigates the relationship between speed limits and injury severity. Visuals show how accident frequency and severity vary across speed categories (0–30 km/h up to 110+ km/h). The page highlights the disproportionately high severity at higher speed limits, supporting evidence-based traffic policy decisions.

Key takeaways

Injuries are most frequent on 31–50 km/h and 71–90 km/h roads – mainly urban and peri-urban routes – because of dense traffic and mixed road use.
However, severity rises with speed. At higher speed limits (91+ km/h), there are fewer injuries overall but a much higher share of severe or fatal injuries. The relationship between speed limits and severity is clear but not absolute (Cramér's V = 0.134), as other factors like road design and traffic mix also play a role.

Interestingly, roads with 71–90 km/h limits show the highest relative risk, combining frequent accidents with high severity – making them critical targets for safety interventions.

2.7 Special case: bikes / e-bikes page



Description of the page

This focused analysis isolates bike and e-bike accidents and compares them with accidents involving other vehicles. It analyzes helmet use, injury severity, speed limits and total accident counts, showing that bike and e-bike-related accidents have a higher relative risk of severe accidents compared to other vehicles, and highlights speed environment as the most influential driver of cyclist injury severity.

Key takeaways

Cyclists and e-bike users face significantly greater danger than other road users. Between 2019 and 2023, their severe injury share remained consistently higher (26–29% vs. 17–18% for other vehicles). In 2019, their odds of severe outcomes were 81% higher than others, and this gap persisted through 2023.

Helmet users show slightly higher severity rates (likely reflecting riskier riding conditions, not helmet inefficacy).
Speed limit is the dominant factor for both bikes and e-bikes – each increase in speed limit more than doubles the odds of a severe outcome. This highlights speed environment as the most influential driver of cyclist injury severity.

2.8 Measures

2.8.1 Model measures

These core measures summarize the overall scope, severity, and demographic characteristics of accidents. They include:

- Total People in Accidents and its breakdowns by injury category: Non-severe, Severe, Hospitalized wounded, and Fatal Accidents.
- % Severe Injuries and % Non-severe Accidents to evaluate injury distribution.
- Average Age to capture demographic trends among victims.
- YoY % Change to track year-over-year variations in accident volume and assess temporal developments.

Together, these measures form the quantitative backbone for the Executive, Temporal, and User Profile Analysis pages.

2.8.2 Special case measures

This set focuses on bicycles and e-bikes as a distinct analytical category.
It quantifies both exposure and risk through measures such as:

- % of Bikes /e-Bikes in Accidents and % by Injury Severity,
- % Severe Bike and e-Bike Injuries and % Severe Other Vehicle Injuries,
- Total Bike and e-Bike Accidents, Total e-Bikes, and Total Bikes and e-Bikes in Severe Accidents,
- Risk Ratio (Bike and e-Bike vs Other Vehicles) – a comparative metric to assess relative injury risk.

These measures enable deeper insight into micro-mobility safety trends, feeding the Special Case: Bikes/e-Bikes page.

2.8.3 Speedband measures

These measures support the Speed Limit Analysis page by evaluating the relationship between speed categories and accident severity.
They include:

- CI Lower, CI Upper, and CI Offset, representing confidence interval boundaries for the percentage of severe injuries per speed band — used to visualize data reliability and variation across limits.

2.9 Interactions

To ensure interpretability and methodological consistency, we systematically reviewed the interaction behavior between slicers, visuals, and KPIs on each dashboard page (Power BI: Format → Edit interactions). For every page, we tested how each element responded to (i) slicer selections and (ii) click selections on charts (bars, segments, map regions), and then configured the most appropriate interaction type: Filter, Highlight, or None.

Our default pattern was Filter. Slicers and the main navigational visuals apply row-level filtering across the page so that KPIs and comparison charts reflect the same subset of the data. This reduces cognitive load, prevents contradictory states between visuals, and keeps absolute counts and percentages coherent. We deliberately avoided cascading highlight effects when they could produce partial totals or visually emphasize subgroups without actually restricting the underlying denominator.

We used None in cases where cross-filtering would be confusing or computationally expensive. Examples include reference KPIs designed to remain stable (e.g., fixed baselines or targets), explanatory cards (definitions, metadata), or summary visuals intended to show the global context regardless of local selections. Disabling interactions for these elements prevents accidental re-slicing and preserves their explanatory role.

Highlight was applied selectively where it adds explanatory value by showing proportional contributions without altering totals. A representative case is the donut chart “Share of people by injury severity” on the Executive Summary: clicking a user category (e.g., Drivers) highlights only that subgroup within each donut segment while preserving the overall distribution. This makes the intra-category composition visible (“where within the whole”) and avoids changing denominators as a full Filter would. We used this pattern only on visuals where comparative context is essential and the retained totals reduce the risk of misinterpretation.

3. Interpretation of results

Our analysis examined how injury severity of road accidents in France varies by time, user profile, geography, and road speed limit. As a special case we focused on cyclists and e-bike users with respect to safety equipment (helmet use) and the posted speed limit.

Temporal patterns show that severe injuries are most likely during the early morning hours (1–4 AM) and on weekends, especially Sundays, while overall severity has slightly declined from 2019 to 2023.

Age influences severity – older adults and teenagers face much higher risks of serious or fatal injuries than other groups.

Geographically, urban areas (i.e., Paris region) record the most injuries but the lowest severity rates, whereas rural and overseas departments show much higher proportions of hospitalized or fatal outcomes. This highlights disparities linked to speed, infrastructure, and emergency response times.

Severity also increases with speed limits: while most accidents occur on moderate-speed urban roads (31–90 km/h), high-speed roads (91+ km/h) produce a far higher share of severe and fatal cases.

Cyclists and e-bike riders remain particularly vulnerable, with 27% of severe injuries, compared to around 18% for other vehicles. Posted speed limit is the most consistent predictor of injury severity for these users.

After exploring these associations through Cramér's V and logistic regression, we chose visuals to emphasize time-of-day, weekend vs. weekday, speed limits, and regional disparities, which better communicated the real drivers of injury severity.

Line charts over years, months, weekdays and hours effectively reveal temporal risk patterns at a glance.
As the speed limit increases, the likelihood of a severe injury also rises. This clear upward trend is showcased in a line graph showing how the share of people in severe accidents changes across speed limits and includes the model prediction from a binomial generalized linear model.
Pie charts comparing departments visually emphasize contrasts between urban and rural regions. An interactive map with hover details provides intuitive spatial insights.
Interactive filtering (e.g., by age, speed limit, vehicle type, and department) allows us to uncover relationships and makes it easy to explore how risk changes for cyclists versus other users, by speed limit and between urban and rural areas.

These visualization choices transform raw data into a coherent narrative, helping audiences quickly grasp where, when, and for whom injuries are most severe, while guiding policy recommendations for speed management, infrastructure improvements, and cyclist protection.

Part C: Conclusion

1. Final reflections on the project

1.1 Summary of the project

This project provided a comprehensive analysis of road traffic accidents in France between 2019 and 2023, integrating data auditing, statistical analysis, and visualization to understand the key factors influencing accident severity. Through the systematic exploration of temporal, demographic, geographic, and environmental dimensions, we identified clear and actionable insights into when, where, and for whom road accidents are most severe.

Our findings reveal that injury severity is not random but associated with identifiable patterns. Temporal trends highlight that weekends and early morning hours present the greatest risk, likely linked to reduced visibility, fatigue, or impaired driving. The influence of age is significant, with both the youngest and oldest road users facing higher probabilities of severe injury. Geographic disparities show that while urban areas report the highest number of accidents, rural and overseas departments experience more severe outcomes, reflecting differences in speed environments, infrastructure quality, and access to emergency services.

Speed limit emerged as the most consistent and powerful predictor of severity: as speed limit increases, so does the likelihood of serious or fatal outcomes. This relationship is particularly critical for cyclists and e-bike users, who are disproportionately represented among severe cases. Their vulnerability underscores the urgent need for protective measures such as dedicated cycling lanes, stricter speed regulation in mixed-traffic areas, and awareness campaigns promoting helmet use.

The use of Python for data cleaning, visualization, and statistical testing (including Chi-square, Cramer's V, and logistic regression) enabled a robust and transparent analytical process. The resulting Power BI dashboard further enhances this work by providing an interactive, intuitive platform for stakeholders to explore the data dynamically and identify targeted interventions.

In conclusion, this project demonstrates how data-driven insights can inform road safety policy and urban planning. Reducing speed limits in high-risk areas, improving rural infrastructure, and strengthening protection for vulnerable users such as cyclists could collectively reduce the severity and frequency of road traffic injuries. Continued monitoring, supported by interactive data tools, will be essential for measuring the impact of such interventions and guiding future strategies toward safer roads across France.

1.2 Difficulties encountered during the project

A first difficulty concerned keys and joins. Only the characteristics table exposes a clear primary key (num_acc). The users, vehicles, and places tables do not provide unique identifiers, which initially led to merge ambiguities and row explosion. We resolved this by anchoring the model on the users table (one row per participant) and using a composite key for the user–vehicle join (num_acc + id_vehicule + num_veh). For the places file - where several rows can exist per accident - we aggregated multiple rows to one row per num_acc (pipe-separated distinct values) before merging, thus preserving cardinalities.

A second challenge was data quality and undocumented codes. Several variables contained missing values, "-1 / not specified" codes, or additional categories not listed in the public documentation. Rather than rely on the codebook, we built a checkcolumn() helper that lists all distinct values with counts and percentages. Using this evidence, we decided variable by variable whether to keep or drop a column and how to handle NaN/-1. As a rule of thumb, a 5% missingness threshold guided inclusion decisions. Specific treatments—such as grouping injury_severity = -1 into "Non-severe", hiding year_of_birth = -1 from age buckets, and handling speed_limit = -1 consistently—are documented in the preprocessing section.

The concatenated Places file (2019 - 2023; before aggregation to one row per accident) contains 289264 rows of data with 7971 rows of unspecified values in the speed limit column (coded with -1). The unspecified values were not included in the analyses or graphs as previously outlined, as they represent 2.8% missing data. However, the speed limit column also contained errors: road speed limits of 0 (n = 1), 1 (n = 67), 2 (n = 47), 3 (n = 9), 4 (n = 4), 5 (n = 106), 6 (n = 27), 7 (n = 3), 8 (n = 2), 9 (n = 1), 10 (n = 486), 12 (n = 2), 23 (n = 1), 31 (n = 1), 42 (n = 1), 180 (n = 1), 300 (n = 7), 500 (n = 41), 501 (n = 1), 502 (n = 1), 520 (n = 1), 560 (n = 1), 600 (n = 1), 700 (n = 4), 770 (n = 1), 800 (n = 1), 900 (n = 4), 901 (n = 1). We were not sure what to do with these 823 unreliable values, as they are not outliers nor missing values, but most likely errors on the part of data entry. For lack of experience and a consistent way to handle these errors, they were left in the dataset and included in the graphs and analyses. We considered the effect size small, especially since we did not use the actual speed limits for analyses (such as for calculating a mean speed limit) but divided them into speed buckets, for example, the speed limits of 900 would add four more values to the speed bucket of "110 + " instead of inflating a mean.

Collaboration in Google Colab introduced concurrency issues. Simultaneous edits occasionally caused execution conflicts and overwrites. We adopted a simple protocol: only one person edited a notebook at a time, with clear hand-offs in chat, frequent checkpoints (file copies with timestamps), and small, modular notebooks to reduce the blast radius of changes.

Power BI collaboration was another constraint. We chose Power BI for its functionality, but the free edition is not built for multi-user editing. To mitigate this, we moved as much work as possible upstream: preparing fact and dimension tables in Python, fixing types and buckets, and exporting lean CSVs. Dashboard construction was then done in joint working sessions on a single Windows machine (best hardware), with the PBIX and source files shared after each session so teammates could validate numbers, test interactions, and prepare next-session visuals.

Finally, intermittent sign-in prompts in Power BI Desktop (requesting a professional account) repeatedly interrupted the build process, especially during longer sessions. While this did not affect the analytical results, it increased development time and cognitive load. We worked around it with frequent saves, periodic restarts, and short, focused build cycles to minimize disruption.

1.3 Continuation of the project

1.3.1 Areas for improvement and future work

This project lasted just under 60 days, which at times required us to make certain compromises and cut corners to meet our deadlines. Had we had more time or resources (or started the project with the knowledge we have now), we would have included more of the available factors, such as 'weather conditions', which had very few missing values. A number of similar columns could have provided a more complete picture of the circumstances contributing to road accidents in France.

In Power BI, additional time and access to a paid version would have enabled us to further enhance our storytelling and user experience by adding animations, navigation buttons, and other interactive design elements. We also identified several technical improvements to increase performance and analytical depth:

- Using Power BI parameters or bookmarks to streamline navigation between themes and pages.

- Implementing incremental data refresh for better scalability and reduced loading time.

- Refining data model relationships by adding new dimension tables (i.e. Dim_VehicleType or Dim_RoadCondition) to simplify filtering and improve data organization.

- Creating additional DAX measures to provide automatic comparisons, such as accident rate per 100,000 inhabitants or severity ratio per department to make the analysis more insightful and interactive.

Finally from a data preparation perspective, some columns (i.e. Speed Limit) contained inconsistent or erroneous data, while others changed in structure over time, making them difficult to integrate. With more time, we would have devoted additional effort to data cleaning, addressing missing values, duplicates, and inconsistencies to ensure the most accurate representation possible.

1.3.2 Key insight and final reflections

Through our analysis and the development of the Power BI dashboard, we were able to go beyond simply counting the number of road accidents to explore how different factors relate to both injury severity and number of people in accidents. By combining descriptive statistics, visual analytics, and statistical tests, we gained a deeper understanding of which variables truly have an influence and to what extent.

Our dashboard, which we structured across five topics (temporal, user profile, speedband, location, and special cases: bikes and e-bikes) allowed us multiple perspectives to analyze this data. For example, we could visualize the most affected user groups by age and gender, assess how speed limits correlate with severity, and, through the use of slicers, observe how these relationships change by year, department, or vehicle category.

Our use of severity percentages and ratios, along with statistical tests such as Cramer's V, Chi-square, and Logistic Regression, provided meaningful insights into our selected factors that most strongly affect road accident outcomes in France.

Overall, this project not only provided a solid technical exercise in data cleaning, modeling, and visualization but also offered meaningful insights into the dynamics of road accidents in France between 2019 and 2023. Our findings highlight how structured data analysis and interactive visualization can transform complex datasets into an accessible and coherent understanding of real-world phenomena.

2. Bibliography

Ministry of the Interior (2025) Annual databases of road traffic injuries – Years 2005 to 2024. [Dataset]. data.gouv.fr. Last updated 21 October 2025. Available at: https://www.data.gouv.fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2024 [Accessed 23 October 2025].

ONISR (2025) 2024 Road Safety Review. [Online]. Published 28 May 2025. Updated 12 September 2025. Available at: https://www.onisr.securite-routiere.gouv.fr/etat-de-linsecurite-routiere/bilans-annuels-de-la-securite-routiere/bilan-2024-de-la-securite-routiere [Accessed 23 October 2025].