**Historic Road Accidents in France – A Study**

Carlos Natale

Stephen Waller

Ehsan Jafari

**Report 1 : Exploration, Data Visualisation and Data Pre-Processing**

**Context**

For this project, we are tasked with conducting a detailed analysis of the history of road accidents in France using data provided by the French government through the Ministry of the Interior and Overseas Territories.

The objective is to analyze the available data, clean and format it technically, and to understand any correlation between data items that could provide insight and statistical significance of the data objects. This can lead to improved predictions of the likelihood of road accidents and their severity. We will also utilize advanced data analytics techniques, such as machine learning, to build predictive models and derive actionable insights.

In the long term, this analysis could bring economic benefits to both the population of France and the French government. The reduction of accidents or their severity would have far-reaching benefits due to the socio-economic impact of road accidents – such as road closures, injuries, or even death. Accident prevention will help reduce healthcare costs, alleviate the financial burden on victims and their families, and minimize the economic impact of traffic disruptions. Furthermore, enhanced road safety will lead to increased productivity and economic efficiency.

From a scientific standpoint, this project contributes to road safety research by providing a data-driven approach to understanding accident patterns and their causes. It supports the development of predictive models and safety interventions based on empirical evidence, enhancing our ability to prevent accidents and save lives.

**Objectives**

The main objective of this project/report is to have the data analyzed, processed, cleaned and combined as required, in order to have a singular dataset that can be used for complex statistical analysis and modeling and create predictive models to forecast accident hotspots and severity.

All members of the group have differing levels of expertise in data science, but we are all working towards improving our knowledge as part of the training with DataScientest.

**Understanding and manipulation of data**

**Framework**

The data that is used for this project is sourced from the Ministry of the Interior and Overseas territories, and is freely published by the French government via the following webpage. [*Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2022 -* ***data.gouv.fr***](https://www.data.gouv.fr/en/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2022/)

The data on this site contains details of road accidents from 2005 to 2022 – for the purposes of this project, as a team, we have agreed to focus on the data sets from 2019-2022, as 4 years worth of recent data would be able to give a valuable enough insight, and reducing the risk of results being skewed by using more historic data where the severity of an accident may be affected by differing standards in vehicle safety etc.

The datasets themselves consisted of 4 .csv files per year - “usagers”, “vehicules”, “lieux” and “carcteristiques” - users, vehicles, locations, and characteristics related to an accident.

Considering the years 2019 to 2022, the volume of records for each dataset is:

* characteristics: **218,404** rows
* locations: **218404** rows
* vehicles: **373,584** rows
* users: **494,182** rows

## **Relevance**

As a team we have agreed on the “grav” (Severity) field as the target variable for this project as a whole, this field describes the severity of injuries as a result of this accident. Ranging from “uninjured” to “Fatal” accidents, this target was chosen as the severity of an accident can be directly linked to a number of factors in the datasets, such as location and speed.

There are limitations to the data, as this data is only present where an accident has been logged by a law enforcement unit (Police etc.), so any accidents that have occurred without a report being written, or attendance of law enforcement, would not be present on the dataset(s), so a complete picture may not be possible.

**Relevant Variables:** The fields were selected according to the items described in “Data Cleaning and Processing”.

|  |  |
| --- | --- |
| **Category** | **Columns** |
| **Identifiers** | AccID, vehicleID, num\_veh |
| **Temporal Variables** | birth\_year, day, month, year, time |
| **Spatial Variables** | lat, long |
| **Accident Characteristics Variables** | collision\_type, initial\_impact\_point, fixed\_obstacle, mobile\_obstacle, accident\_situation |
| **Environmental Variables** | lum, atm\_condition |
| **Demographic Variables** | user\_category, gender, age |
| **Vehicle-Specific Variables** | vehicle\_category, motor |
| **Human Factors Variables** | maximum\_speed, manv, seat, reason\_travel, safety\_equipment1 |
| **Target Variable** | **gravity** |
| **Road Characteristics Variables** | route\_category, traffic\_regime, total\_number\_lanes, upstream\_terminal\_number, distance\_upstream\_terminal, plan, surface\_condition, infra, traffic\_direction, reserved\_lane\_code, longitudinal\_profile |

**Target Variable:**

The primary target variable is the severity of injuries sustained, categorized into:

* Indemne (Uninjured)
* Tué (Fatal)
* Blessé hospitalisé (Hospitalized injury)
* Blessé léger (Minor injury)

**Dataset Features:**

The dataset provides a comprehensive view of each accident, including:

* Detailed accident descriptions (time, location, conditions).
* Demographic information of involved parties.
* Vehicle details.

## 

## **Pre-processing and feature engineering**

**Data Cleaning and Processing**

Extensive data cleaning and processing were necessary to prepare the dataset for analysis. The treatment process involved key steps:

* **Merging Datasets**

The preprocessing began by merging several CSV files across four distinct datasets—accidents, locations, users, and vehicles—collected from 2019 to 2022. These files were loaded and consolidated into a single dataset, ensuring all relevant data related to each accident was included for comprehensive analysis.

* **Standardizing Column Names**

To facilitate readability and collaboration, especially within an English-speaking team, all column names were translated from French to English. This ensured consistency and clarity across the different datasets, enabling easier understanding and processing of the data.

* **Handling Missing and Outlier Values & Irrelevant Fields**

Replacement of Specific Values: In cases where the dataset contained `1` to indicate "Not specified" values (such as in the `reason\_travel` column), these were replaced with more meaningful values like `'0'`, converting "Not specified" to "Unknown." Additionally, all occurrences of `1` across the dataset were replaced with `NaN` to standardize missing data representation.

Column Removal: The dataset contained some columns deemed irrelevant for the analysis, such as `id\_usager`. These columns were removed. Furthermore, columns with more than 30% missing values were dropped to ensure the integrity of the data and avoid bias caused by incomplete data.

* **Duplicate and Outlier Removal**

**Duplicates**: Any duplicate records in the dataset were identified and removed to ensure that each accident was represented only once.

**Outliers**: Extreme or implausible values were also treated. For example:

**Speed**: Values lower than 5 km/h or greater than 125 km/h were removed from the `maximum\_speed` column as they were considered unrealistic.

**Age**: Based on statistical testing, values below 0 and above 97 were removed as outliers in the `age` column.

**Geographical Codes:** Outlier values in certain geographical columns, such as `dep\_code`, were also removed based on predefined rules to ensure data consistency.

* **Latitude and Longitude Conversion**

To maintain consistency in geographic coordinates, all commas (`','`) used as decimal points in the latitude and longitude columns were converted to periods (`'.'`). This ensured proper handling of geographic data and prevented errors during subsequent geospatial analysis.

* **Missing Value Imputation**

Missing values in critical fields were addressed through imputation. Rather than simply dropping rows with missing values, imputation was performed based on the distribution of existing values in each column. This allowed the dataset to maintain its integrity without introducing significant bias or reducing the overall size of the dataset.

* **Deletion of Redundant Location Fields**

While the dataset contained multiple fields related to accident location, only the latitude and longitude columns were retained. Other location fields were removed to avoid redundancy and to focus on the most critical geographical information for the analysis.

This preprocessing phase ensured that the data was clean, consistent, and ready for use in subsequent analytical and modeling tasks aimed at predicting accident severity and identifying key factors contributing to road accidents in France.

|  |  |
| --- | --- |
| **Raw Data** | **Processed Data** |
|  |  |

We used Ridge regression to shrink coefficients and reduce overfitting to provide insight into how each feature contributes to the target variable. Below are the regression coefficients for each feature in the model.

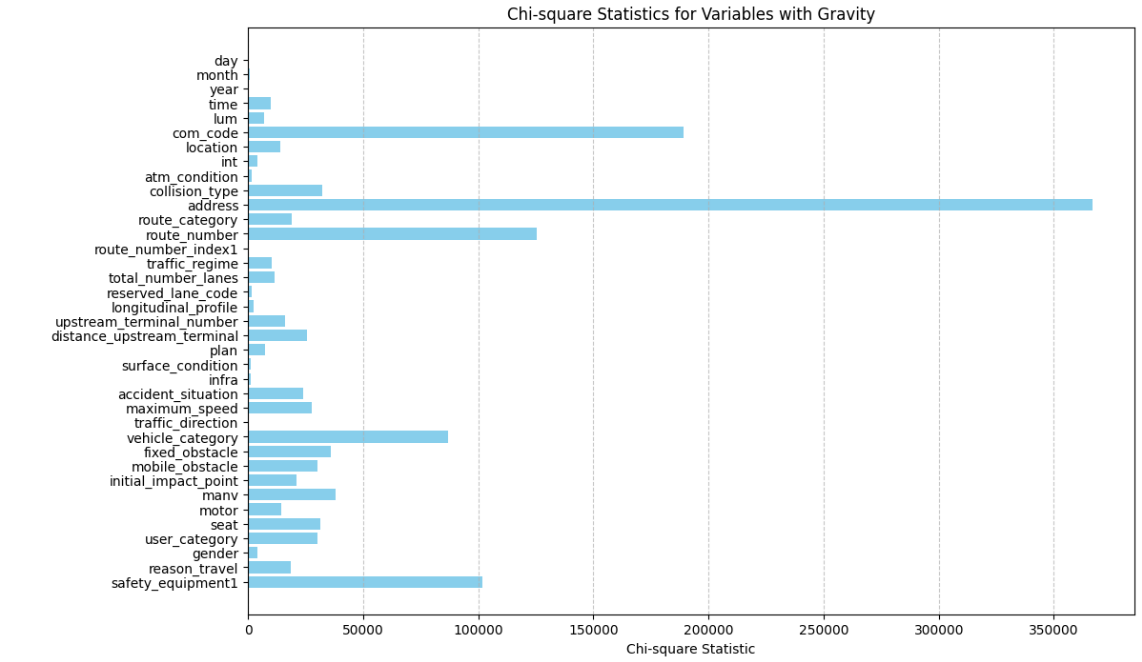
* **Positive Coefficients**: Features with positive coefficients increase the predicted target value (e.g., accident severity) as their values increase.
  + **user\_category (0.366)**: The type of user (driver, passenger, pedestrian) has the strongest positive influence on the target.
  + **motor (0.208)** and **gender (0.196)**: These factors also show significant positive contributions, indicating their importance in predicting accident severity.
  + **maximum\_speed (0.061)**: Higher speed limits slightly increase the predicted severity of accidents.
  + **seat (0.046)**, **fixed\_obstacle (0.044)**, and **plan (0.042)**: These variables contribute positively but less substantially, possibly linked to road conditions or safety measures.
* **Negative Coefficients**: Features with negative coefficients decrease the predicted target value.
  + **mobile\_obstacle (-0.094)**: This has the most significant negative impact, suggesting that mobile obstacles reduce the severity of accidents.
  + **collision\_type (-0.055)**: Certain collision types are associated with lower severity.
  + **reserved\_lane\_code (-0.044)**: Reserved lanes may reduce accident severity due to better traffic management.
  + **age (-0.012)** and **year (-0.016)**: Older individuals and more recent years correlate with decreased accident severity.
* **Near-Zero Coefficients**: Features with coefficients near zero have minimal impact on the target variable.
  + **time (-9.73e-10)** and **upstream\_terminal\_number (-6.92e-05)**: These features are insignificant in predicting accident severity.

In the next step, we will analyze the variables using the correlation **matrix methods, the mutual information (MI) scores, and the feature importance scores of a Random Forest model** to confirm or identify the most significant predictors and understand the relationships between the variables.

After all the pre-Processing, the files have been merged and the total number of records in the Dataset is **447,670** rows.

## **Visualizations and Statistics**

Relationships between our target variable of “grav” (Severity) have been identified via various methods of analysis, for example chi-square analysis (see graph below) has been done on the overall dataset against the “grav” field showing that there are key variables that are significant to be used to model against in the future – and example of this would be the “route number” which gives details on specific roads / road types – which have a correlation with the severity of an accident. As part of modeling, this would be a good data item to use, to see how strong the correlation between road types and severity – a hypothesis of this would be more rural roads would have a higher severity of accident.



**Chart 1 – Number of accidents by Accident Severity**

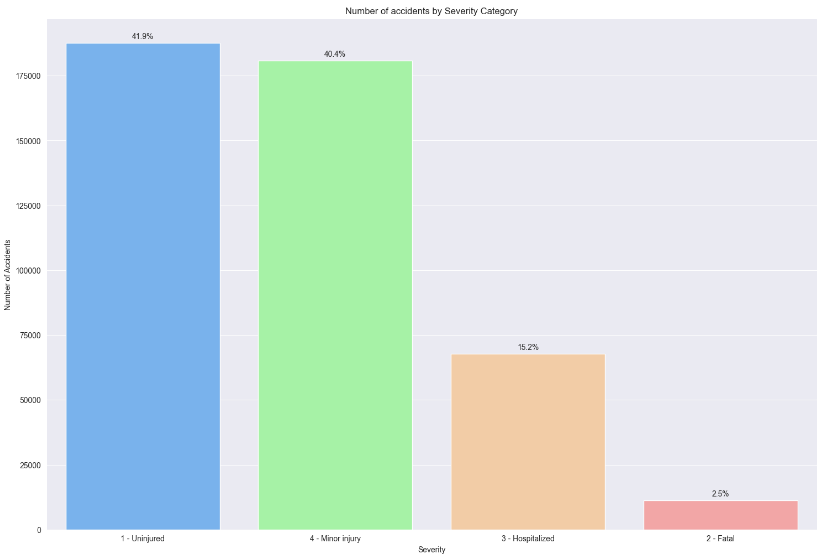


Chart 1 illustrates that the majority of accidents led to either no injury or only minor injuries, accounting for over 80% of the cases. A smaller proportion of accidents resulted in hospitalization, indicating more severe injuries. Fatal accidents are rare, making up a very small percentage of the total. This distribution suggests that while accidents are frequent, the majority do not result in severe harm. Only a small fraction of accidents lead to hospitalization or death.

**Chart 2 – Accident Severity distribution by Year**

|  |  |
| --- | --- |
|  |  |

Chart 2 illustrates a consistent pattern where most accidents result in either no injury or minor injuries, in line with general traffic accident trends. The percentage of fatal accidents is very low but slightly increases over time, which may require closer examination to identify potential causes. The stability in the percentages of hospitalized individuals suggests that while the severity of injuries remains relatively unchanged, the absolute number may vary with the total number of accidents each year.

The combined visualization effectively communicates both the absolute counts and relative percentages of accident severity over the years, providing insights into the nature and trends of traffic accidents during this period.

**Table 1 – Accident Severity distribution by Lightning condition**

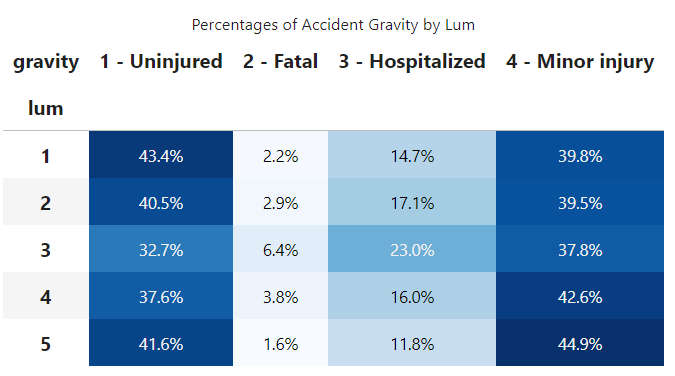


Table 1 indicates a relationship between lighting conditions and the seriousness of accidents. Lum 3 seems to be linked to more severe outcomes (higher fatality and hospitalization rates), which may be linked to poor or challenging lighting conditions (night without public lighting). Lum 1, likely representing optimal lighting (Daylight), shows a higher proportion of uninjured and minor injuries, which aligns with safer driving conditions. Lum 5, while displaying the highest percentage of minor injuries, also has the lowest fatality rate, which could suggest good but not perfect lighting conditions (night with public lighting).

**Chart 3 – Accidents Severity Distribution by Gender**

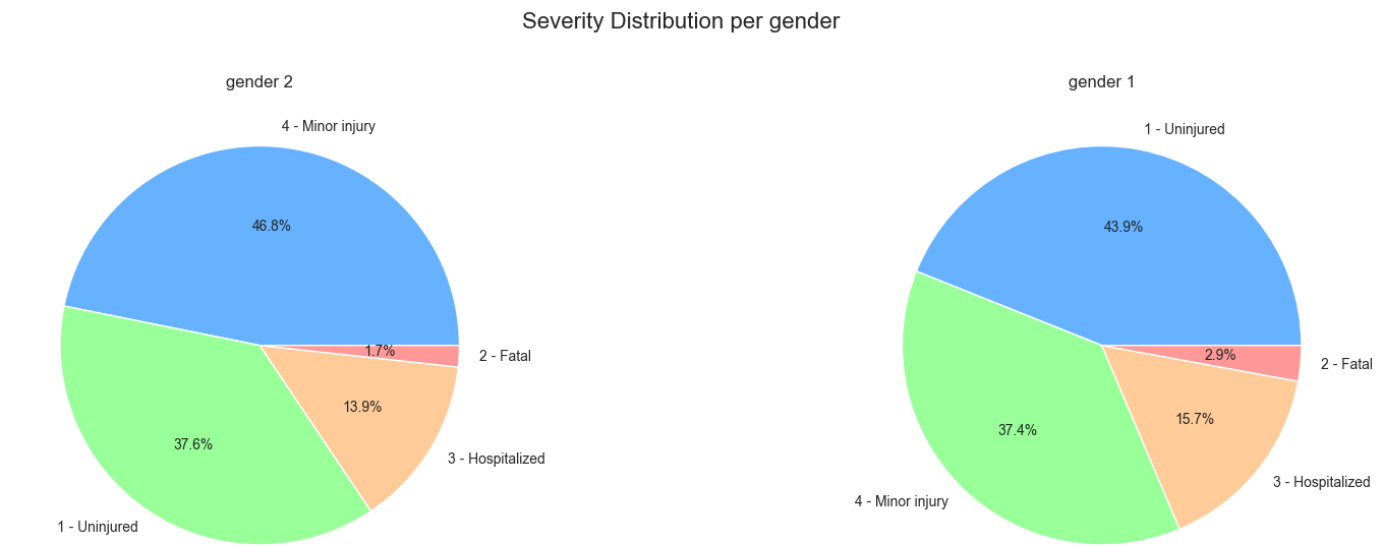


Chart 3 indicates that Gender 2 (Female) has a higher percentage of minor injuries and a lower percentage of fatalities compared to Gender 1 (Male). Gender 1 has a higher percentage of accidents resulting in hospitalization and fatalities, indicating that accidents for this group tend to be more severe. Gender 1 also has a higher distribution of uninjured accidents, while minor injuries are more common in Gender 2. These differences may reflect factors such as driving behavior, risk exposure, or vehicle types associated with each gender.

**Chart 4 – Age Distribution by Accidents Severity**

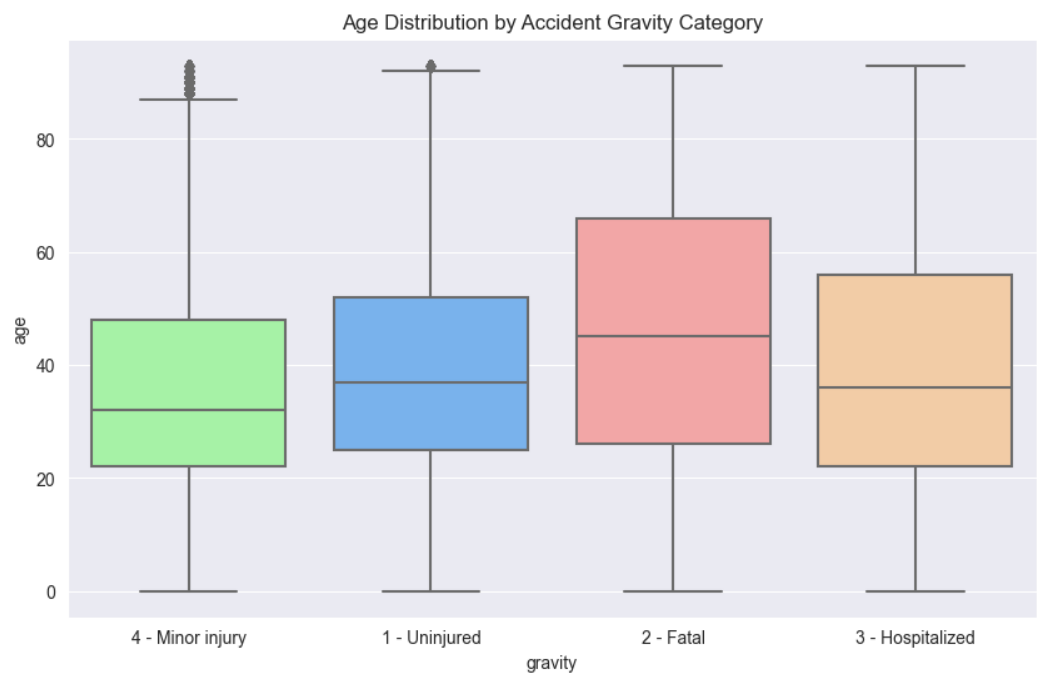


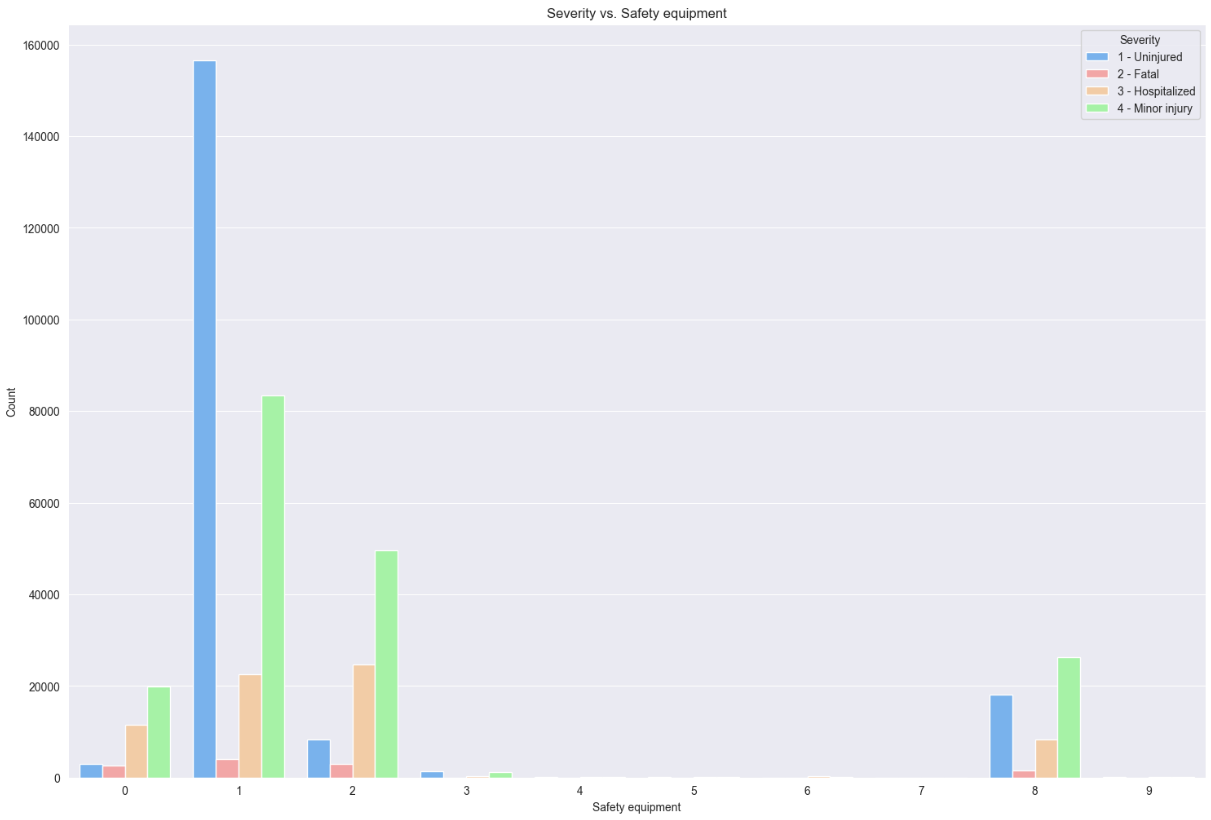
Chart 4 shows that fatal accidents are more common among older individuals, with the highest median age compared to other categories. Uninjured and Minor Injury categories have similar age distributions, with younger median ages. Hospitalization affects a wide age range, with a median age slightly lower than fatal accidents but higher than minor injuries and uninjured cases. The presence of outliers in the Minor Injury and Uninjured categories suggests that accidents affect individuals of all ages, though the severity of outcomes tends to increase with age. This graph suggests the existence of a relationship between age and accident severity, showing that older age groups are more susceptible to severe outcomes such as fatalities.

**Chart 5 – Accident Severity by Age vs. Maximum Speed**



Chart 5 shows that higher speeds consistently result in fatal accidents across all age groups, especially among individuals aged 20-60. Among younger individuals, different speeds indicate a higher risk of severe accidents (fatal or requiring hospitalization) when speed is a factor. As individuals get older, the maximum speed decreases, and the severity of accidents tends to level off, indicating that speed has less influence on accident outcomes. This chart highlights the strong connection between higher speeds and the severity of accidents, especially fatal ones, across most age groups. It emphasizes the crucial role of speed management in reducing severe outcomes, particularly for younger and middle-aged drivers.

**Chart 6 - Accidents Severity Distribution by Safety Equipment**

****

In Chart 6, Category 1: This is the most effective safety equipment, with the highest rate of uninjured individuals, making it the best choice for preventing injuries in accidents. Categories 0, 2, 5, 6, and 9 show varying levels of effectiveness. The chart highlights the importance of safety equipment in reducing the severity of accidents, though it also shows that certain types of equipment (or lack thereof) are associated with a higher incidence of minor injuries and hospitalizations.

**Chart 7 - Accidents Severity Distribution by Seat in accident moment**

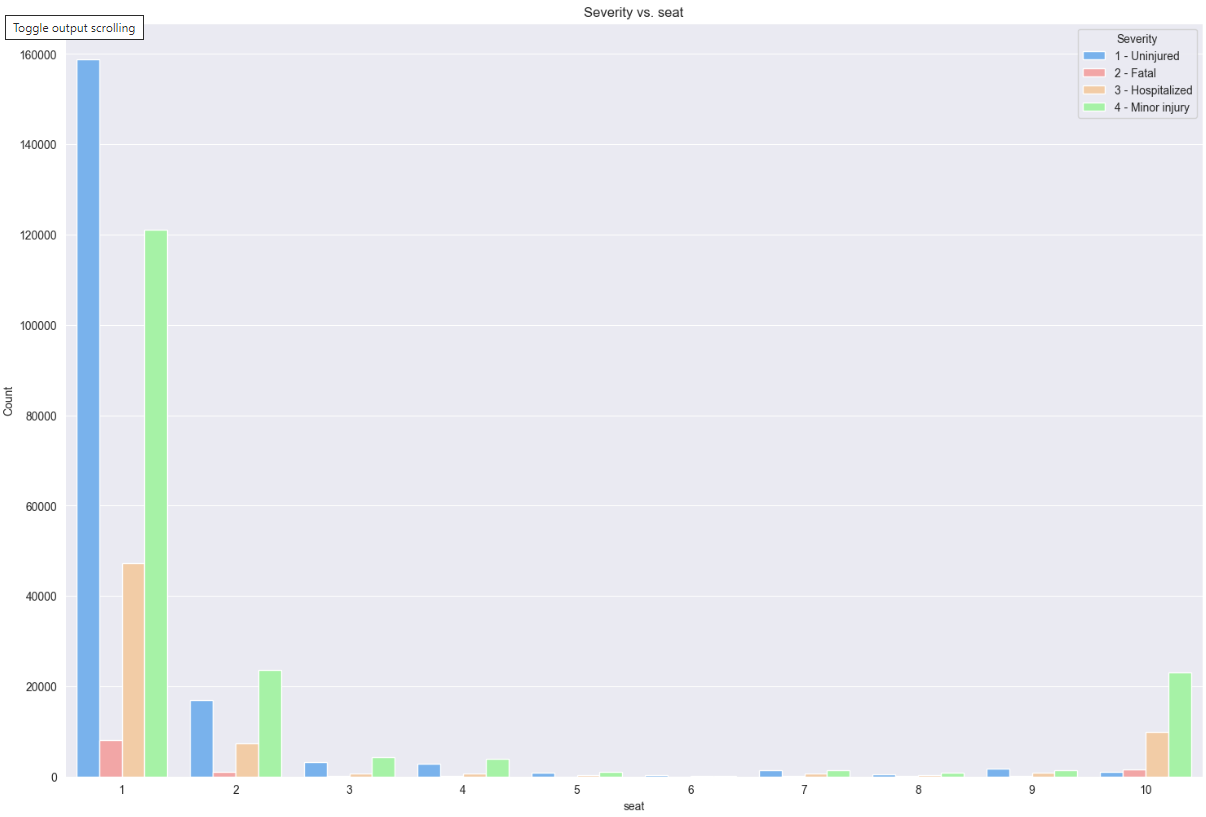
****

Chart 7 indicates that Seats 0 and 1, likely the driver and front passenger seats, have the highest number of accidents. Most of these accidents resulted in no injuries or minor injuries, but there were also a significant number of hospitalizations and some fatalities. Seat 2, possibly a rear seat, shows a more balanced distribution of accident severity outcomes. Seats 3 to 10 had fewer recorded accidents, suggesting that these seats might be less occupied or less prone to severe outcomes. Seat 10 stands out slightly for minor injuries. This chart demonstrates the distribution of accident outcomes across different seat positions, indicating that the front seats are associated with a higher exposure to both minor and severe accidents. This information can be useful in understanding the risks associated with different seat positions and can help inform vehicle safety designs.

**Chart 8 - Accidents Severity Distribution by User Category involved in accident moment**

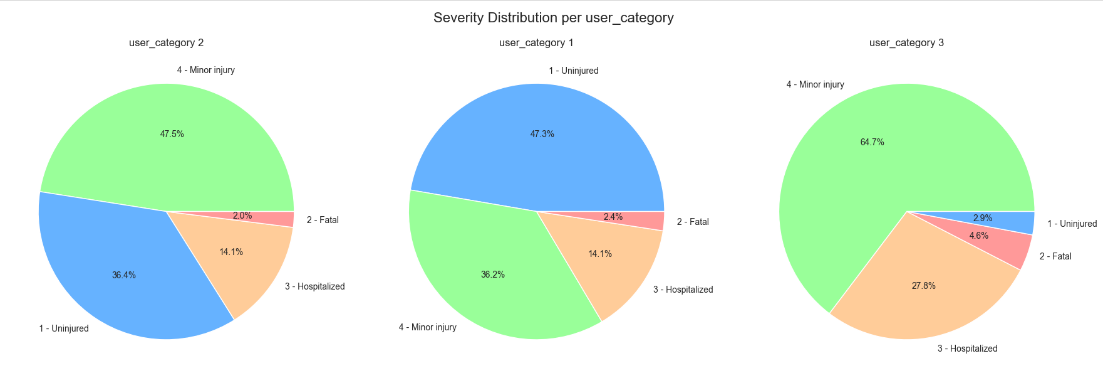


Chart 8, User Category 1 (Driver) appears to be safer, with a higher percentage of uninjured outcomes and lower fatality rates. User Category 2 (Passenger) shows a slightly higher risk compared to User Category 1, with more minor injuries and a similar hospitalization rate. User Category 3 (Pedestrian) seems to be the most vulnerable, with a higher rate of minor injuries, hospitalizations, and fatalities. This suggests that this group, possibly pedestrians or cyclists, faces greater risks in accidents. This visualization highlights the varying risk profiles across different user categories, indicating that safety measures may need to be tailored to better protect more vulnerable groups, such as those in User Category 3.

## 

## **Next stage of modeling**

Now that the final variables have been decided upon, and there is a “clean” (but still raw) dataset that encompasses the variable that are best suited to affect the “grav” (severity) target variable, the next stage of or project is to begin the modeling task to see how effective these variables are in predicting our target.

As part of this modeling process there will be further feature engineering covered - to create more variables - for example a creation of an overall date/time field that covers the accident, as well as averages of these - this feature would then be used as part of historic/date-time modeling, likely using the ARIMA method - this could provide insight as part of the conclusion of this project to see if the severity of accidents are changing over time, as well as the average number in a given time-frame. This would allow predictions into the future, with (hopefully) a reasonable degree of accuracy around possible numbers of accidents, and their predicted severity.

Other field amendments would be amending the categorical data we have in our dataset to prepare for predictive modeling, methods being considered is Label Encoding (setting categories to a unique integer) or Frequency Encoding where the frequency of the data is important for the next stage of modeling. In addition, Normalization/Standardization and Data Balancing techniques would be applied.

Initial use of a standard test/train split would be done, then this outcome would be applied to various modeling techniques available via Python libraries, such as Gradient Boosting, Decision Trees, and Random Forests. This will enable a further prediction of the target “severity” based upon changing variables that feed in, with a high degree of accuracy.

The outcomes of these will be tuned as appropriate, and then re-run to give the most accurate predictions possible, using the most of the data we have.

## 

## **Conclusion**

This study provides a comprehensive analysis of road accidents in France using recent data from 2019 to 2022. The primary objective was to explore, clean, and preprocess various datasets related to accidents, vehicles, users, and locations to understand the factors influencing accident severity and eventually develop predictive machine learning models that could forecast accident severity.

Extensive data cleaning and preprocessing were crucial steps in this analysis. This involved merging multiple datasets, standardizing column names, handling missing and outlier values, and removing irrelevant fields. Key preprocessing steps included replacing inconsistent values, eliminating duplicates, and converting geographic coordinates to a consistent format.

Our analysis identified significant correlations between accident severity and various factors, such as road types, lighting conditions, demographics, and safety equipment use. Visualization techniques helped uncover patterns, showing, for instance, that fatal accidents are more likely on certain road types and under poor lighting conditions, and that older individuals are more susceptible to severe accident outcomes.

The next phase of the project will involve advanced modeling using machine learning techniques, such as Gradient Boosting, Decision Trees, and Random Forests. Further feature engineering will be conducted, including encoding categorical data to enhance model performance. These models will allow us to predict accident severity with a high degree of accuracy, providing valuable insights for policymakers to implement safety measures aimed at reducing accidents and their associated socio-economic impacts.

Through this data-driven approach, the project aims to contribute significantly to road safety research, offering predictive insights that could help save lives and improve overall traffic safety in France.

**Report 2: Data Processing and ML Models Evaluation**

**Data Preprocessing**

The goal of this preprocessing step is to prepare the data for machine learning modeling by handling categorical variables, standardizing numerical variables, and splitting the data into training and testing sets. Below is a detailed breakdown of the preprocessing steps applied to the dataset.

**Data Loading**

The dataset is loaded from a CSV file using pandas.read\_csv.

A copy of the original dataset is made to ensure that any changes do not affect the original data.

**Dropping Unnecessary Columns**

Several columns that are not directly relevant for the analysis were dropped:

* **AccID**: Likely an accident ID that is a unique identifier, not contributing to the prediction.
* **birth\_year**: Likely used to calculate the age, which is already included.
* **vehicleID**: A unique identifier that doesn't add predictive value.
* **num\_veh**: Number of vehicles, possibly not critical for this analysis.

This ensures that only relevant features are included, helping avoid overfitting or data redundancy.

**Handling Time Related Columns**

The columns related to time (time, day, month, and year) were converted to a numerical format (float64). This step makes it easier to apply numerical transformations like scaling and allows for these features to be used in a model that expects numerical input.

**Feature Selection**

A set of 34 features (both categorical and numerical) were selected to be used as input variables (independent variables). The target variable, gravity, represents the outcome we aim to predict.

**Selected features include:**

* Environmental factors (lum, atm\_condition, surface\_condition),
* Road and traffic related factors (route\_category, traffic\_regime, plan),
* Vehicle related features (vehicle\_category, maximum\_speed, motor),
* Driver or user related attributes (age, gender, reason\_travel, safety\_equipment1).

**OneHot Encoding for Categorical Variables**

Categorical features cannot be directly used in many machine learning algorithms, so they were encoded using OneHot Encoding. This technique transforms each categorical feature into multiple binary columns (0 or 1), where each column represents a possible category.

The drop\_first=True parameter avoided the "dummy variable trap" by dropping one category from each feature. This ensures that the resulting binary columns do not introduce multicollinearity into the dataset.

**Impact on the shape of the DataFrame:**

The original categorical features were replaced by many binary-encoded features, increasing the number of columns in the dataset.

After encoding, the number of features increased because each category was converted into its binary column. The original 34 features likely resulted in many more columns due to the encoding process. However, the exact number of new columns depends on the number of unique categories in the categorical variables.

**Train Test Split**

The dataset was split into training and testing sets using an 8020 split:

* 80% of the data is used for training the model.
* 20% is reserved for testing the model’s performance on unseen data.

This split was done using train\_test\_split from sklearn, with a fixed random\_state of 42 for reproducibility.

**Standardization of Numerical Variables**

The numerical features were scaled using StandardScaler from sklearn, standardizing the data by subtracting the mean and scaling to unit variance.

This ensures that all numerical features are on the same scale, which is essential for many machine learning algorithms, especially those sensitive to feature magnitudes like linear models and distance-based algorithms.

Only numeric columns were scaled, ensuring that categorical columns (already encoded as binary) were left untouched.

**Data Shapes**

After all the preprocessing steps, the shape of the datasets was as follows:

* Training set (X\_train): The training set has the processed features after encoding and scaling.
* Test set (X\_test): The test set also contains processed features and is used for model evaluation.

The exact dimensions of these sets were printed out to verify the preprocessing steps:

* Shape of X\_train: (358136, 34)
* Shape of X\_test: (89534, 34)

The increase in features after encoding is expected due to the OneHot Encoding applied to categorical variables.

**Conclusion**

The preprocessing steps were performed successfully to prepare the data for machine learning. Using OneHot Encoding for categorical variables ensures that the model can properly interpret categorical data, and standardization of numerical features helps improve model performance. Finally, the train test split ensures that the model will be evaluated on unseen data to measure its effectiveness.