**Project Report**

**Road accidents in France**

**Exploration, data visualization and data pre-processing report**

**Data Scientest**

**Data Scientist Academic Certificate Program**

**Team** **members**:

Falk Kegler  
Alessandro Perani  
Jennifer Pousada Alvarez  
Ahed Abdelky

# Introduction to the project

## Context

The project aims at providing a Machine Learning-powered solution that helps predict the severity of road accidents in France. Data about previous road accidents in France has been collected over many years. This historical data can be used to train and test ML model(s) to spot and learn correlations between different aspects about road accidents and their severity and location.

Those models are then used to predict potential severity of road accidents based on their corresponding characteristics. This helps to create risk zones with dynamic risk scores according to meteorological information and geographical location among other information.

## Objectives

This solution can potentially help reduce amount of high severity road accidents. Generally speaking, such analysis can be useful for different agencies and companies; e.g.:

* Roads and transportation authorities in towns/cities can benefit from such analysis to design and optimize traffic infrastructure for safety, e.g., improving road safety (speed limits, signage, number of lanes, safety lanes, emergency refuge areas, etc.), planning of new roads, scheduling of road maintenance, etc…
* Insurers can use such analysis for processes like risk assessments, pricing, claims handling for Motor insurance
* Car manufacturers can use such analysis to identify areas of improvement for car safety
* Industries and retail that heavily rely on effective supply chain can use such analysis to avoid substantial disruptions in supply chain
* Navigation system providers and autonomous vehicles manufacturers can use such analysis to deliver functionalities like planning trip routes and suggesting safer roads

This all could potentially help save a lot of lives -millions of people sustain severe or fatal injuries each year- and ease heavy financial burdens on people and institutions.

As all team members work at Allianz, each of us has at least some exposure and/or expertise in one or more topics related to or similar to the problem addressed in this project.

At Allianz, understanding and analyzing frequency and severity of insurable risks are a big competitive advantage. Creating classes or segments with different risk scores is a vital step in pricing insurance products and reserving sufficient funds for future claims. Creating risk zones based on geography for example is a vital part of pricing and reserving for NatCat insurance or insurance of similar risks (hazards). At the same time, predicting severity of road accidents is one of the two main actuarial tasks that actuaries try to model for Motor insurance. Predicting severity of road accidents using ML models and data shared about the accident (claims data) is becoming more widely used, this could help to triage claims, automate some claims handling steps, predict large claims common in the case of bodily injury or predict a total loss (i.e., vehicle needs to be replaced).

We are therefore confident that we have enough exposure to be able to work on the project and solve the problem without further business help. Our target is to successfully deliver an excellent solution which we hopefully can offer to our colleagues at Allianz to improving some of their processes.

# Understanding and manipulation of data

## Framework

Historical data used for the purpose of this project consist of data originally collected about each accident occurring in France on a road open to public traffic, involving at least one vehicle and causing at least one victim requiring treatment. The data describing the accidents was entered by a law enforcement agent present at the accident scene in a pre-defined injury accident analysis bulletin form. The collection of those forms constitute the national traffic accidents file known as the “BAAC file”. The file is administered by the National Inter-ministerial Road Safety Observatory “ONISR”. Datasets extracted from the BAAC registry (called the Etalab database) are made publicly available for free on the data.gouv.fr website and will be used for the purpose of this project.

The datasets contain for each accident: location information as reported, information regarding the characteristics of the accident and its location, the vehicles involved and their victims. The datasets are divided into 4 different file categories:

1. Characteristics: include data that describes the general circumstances of the accident
2. Locations: include data about the main location of the accident
3. Vehicles: include data that describe the vehicles involved in the accident
4. Users: include data about the people involved in the accident whether uninjured or victims. Here “Users” will be used to refer to them

One file is provided per category for each year between 2005 and 2022. Total number of files used for this project is therefore: 4 files x 18 years = 72 files. The files are of the type CSV and their sizes varies between 2.7MB and 12.4MB and add up to 373.1 MB.

The data recorded for of any single accident is split in the four different file categories but can be linked to each other using an accident identifier called: “Num\_Acc”. One accident can of course involve several vehicles and users. Vehicles and their occupants can be linked using a vehicle unique identifier called “id\_vehicule”.

Most variables contained in the four previously listed files contain missing values that were not imputed by the law enforcement agents involved. Missing values are either empty cells or replaced by a zero or period.

Data in all files from year 2005 to 2018 is separated by a “,”with the exception of only one file, the Characteristics file for year 2009 uses a tab “\t” as a separator. From 2019 until 2022, the data is separated in all files by a “;”.

The following tables include the complete lists of fields in all 4 file categories with further details about each field:

**Characteristics**:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name of the Attribute** | **Target variable?** | **What does this variable represent?** | **Data type** | **% of missing values** | **All categories (categorical variables with 10 or less categories). Or basic descriptive statistics on distribution for quantitative variables** | **Range min** | **Range max** | **Comments** |
| NUM\_ACC | NO | Accident ID number | INT64 | 0% |  |  |  |  |
| JOUR | NO | Day of Accident | INT64 | 0% |  | 1 | 31 | Mind the Holidays in France |
| MOIS | NO | Month of Accident | INT64 | 0% |  | 1 | 12 | Mind the Holidays in France |
| AN | NO | Year of Accident | INT64 | 0% |  | 5 | 18 |  |
| HRMN | NO | Hour and minute of Accident | INT64, char | 0% |  | 1 | 2359 | Recode: Time of the day, like morning rush-hour |
| LUM | NO | Lighting conditions in which the accident occurred. | INT64 | 0% | 1, 2, 3, 4, 5 | 1 | 5 |  |
| DEP | NO | Department | INT64, char |  |  | 10 | 976 | Attention: Oversea-Departments Recode to be done in inland departments (timed by 10) A couple of values not int ('2B', '2C') |
| COM | NO | Municipality | FLOAT64 | 0% |  | 1 | 938 | Attention: Oversea-Departments |
| AGG | NO | Location | INT64 |  | 1, 2 | 1 | 2 |  |
| INT | NO | Intersection | INT64 |  | -1, 1, 2, 3, 4, 5, 6, 7, 8, 9 | 0 | 9 |  |
| ATM | NO | Atmospheric conditions | FLOAT64 | 0,018% | -1, 1, 2, 3, 4, 5, 6, 7, 8, 9 | 1 | 9 |  |
| COL | NO | Type of collision | FLOAT64 | 0,001% | -1, 1, 2, 3, 4, 5, 6, 7 | 1 | 7 |  |
| ADR | NO | Postal address: variable entered for accidents in urban areas. | OBJECT | 16,7% |  |  |  | Attention: Oversea-Departments |
| GPS | NO | MISSING | OBJECT | 56,4% | M, A, R, G, Y |  |  | Definition missing |
| LAT | NO | Latitude | FLOAT64 | 56,8% |  |  |  | Many NaN, different value criteria in different input files |
| LONG | NO | Longitude | OBJECT | 56,8% |  |  |  | Many NaN, different value criteria in different input files |

**Places**:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name of the Attribute** | **Target variable?** | **What does this variable represent?** | **Data type** | **% of missing values** | **All categories (categorical variables with 10 or less categories). Or basic descriptive statistics on distribution for quantitative variables** | **Range min** | **Range max** | **Comments** |
| NUM\_ACC | NO | Accident ID number | INT64 | 0% |  |  |  |  |
| CATR | NO | Road category | FLOAT64 | 0% | 1, 2, 3, 4, 5, 6, 7, 9 | 1 | 9 |  |
| VOIE | NO | Road number | Object | 9,52% | Numbers are more frequent in comparison to street names |  |  |  |
| V1 | NO | Numerical index of the road number | FLOAT64 | 54,03% | Value can be 0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, -1.0 | 1 | 3 | Many missing values |
| V2 | NO | Alphanumeric road index letter | Object | 95,19% | 74 different values, “A” and “D” most frequent |  |  | Data is mostly missing. very few entries |
| CIRC | NO | Circulation regime | FLOAT64 | 0,13% | 0.0, 1.0, 2.0, 3.0, 4.0, and -1.0, “2.0” by far most frequent | 1 | 4 |  |
| NBV | NO | Total number of traffic lanes | Object | 0,23% | 69 different values | 0 | 99 |  |
| VOSP | NO | Indicates the existence of a reserved lane, regardless of whether or not the accident takes place on this lane. | FLOAT64 | 0,23% | 0.0, 1.0, 2.0, 3.0, -1.0, most frequent is 0.0 | 0 | 3 | 0 is for not applicable |
| PROF | NO | Long profile describes the slope of the road at the location of the accident | FLOAT64 | 0,17% | 0.0, 1.0, 2.0, 3.0, 4.0, -1.0,  most frequent is 1.0 | 1 | 4 |  |
| PR | NO | PR connection | Object | 40,4% | 1413 different values, 0.0 and 0 most frequent | 0 |  | Many missing values |
| PR1 | NO | Distance to this PR (in m) | Object | 40,56% | 3708 different values, 0.0 and 0 most frequent | 0 |  | Many missing values |
| PLAN | NO | Plan layout | FLOAT64 | 0,19% | 0.0, 1.0, 2.0, 3.0, 4.0, -1.0, most frequent is 1,0 | 1 | 4 |  |
| LARTPC | NO | Width of the traffic island if it exists (in m) | Object | 23,29% | 469 different values, 0 is most frequent |  |  |  |
| LARROUT | NO | Width of the roadway allocated to vehicle traffic does not include emergency lanes, traffic islands and parking spaces (in m) | Object | 9,59% | 728 different values, 0 is most frequent |  |  |  |
| SURF | NO | Surface condition | FLOAT64 | 0,16% | 0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, -1.0. Most frequent is 1.0. | 1 | 9 |  |
| INFRA | NO | Development / infrastructure | FLOAT64 | 0,46% | 0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, -1.0. Most frequent is 0.0 | 0 | 9 | 0 is for none (no development or infrastructure) |
| SITU | NO | Accident situation / location | FLOAT64 | 0,42% | 0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0, -1.0. Most frequent is 1.0. | 1 | 8 |  |
| VMA | NO | Maximum speed authorized at the scene and at the time of the accident. | FLOAT64 | 81,44% | 47 different values, 50 most frequent, followed by 80, 30, 70, and 90 | 0 | 900 | Many missing values |
| ENV1 | NO | Proximity to school | FLOAT64 | 19,02% | 0.0, 3.0, 99.0, most frequent is 0.0 | 0 | 99 |  |

**Vehicles**:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name of the Attribute** | **Target variable?** | **What does this variable represent?** | **Data type** | **% of missing values** | **All categories (categorical variables with 10 or less categories). Or basic descriptive statistics on distribution for quantitative variables** | **Range min** | **Range max** | **Comments** |
| NUM\_ACC | NO | Accident ID number | FLOAT64 | 0% |  |  |  |  |
| ID\_VEHICULE | NO | Unique identifier of the vehicle taken for each user occupying this vehicle (including pedestrians who are associated with the vehicles which struck them) - Numerical code |  |  |  |  |  | not in dataset before 2019 |
| NUM\_VEH | NO | Unique identifier of the vehicle taken for each user occupying this vehicle (including pedestrians who are associated with the vehicles which struck them) - Alphanumerical code | Object | 0% |  |  |  |  |
| SENC | NO | Flow direction | FLOAT64 | 0,02% |  | -1 | 3 |  |
| CATV | NO | Vehicle category | INT64 | 0% |  | -1 | 99 |  |
| OBS | NO | Fixed obstacle hit | FLOAT64 | 0,06% |  | -1 | 17 |  |
| OBSM | NO | Moving Obstacle hit | FLOAT64 | 0,05% |  | -1 | 9 |  |
| CHOC | NO | Initial shock point | FLOAT64 | 0,02% |  | -1 | 9 |  |
| MANV | NO | Main maneuver before incident | FLOAT64 | 0,03% |  | -1 | 26 |  |
| MOTOR | NO | Type of vehicle engine |  |  |  | -1 | 6 | not in dataset before 2019 |
| OCCUTC | NO | Number of occupants on public transport | INT64 | 0,00% |  | 0 | 900 |  |

**Users**:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name of the Attribute** | **Target variable?** | **What does this variable represent?** | **Data type** | **% of missing values** | **All categories (categorical variables with 10 or less categories). Or basic descriptive statistics on distribution for quantitative variables** | **Range min** | **Range max** | **Comments** |
| NUM\_ACC | NO | Accident ID number | int64 | 0% |  |  |  |  |
| ID\_USAGER | NO | Unique identifier of the user (including pedestrians who are associated with the vehicles which struck them) – Numerical code. |  |  |  |  |  |  |
| ID\_VEHICULE | NO | Unique identifier of the vehicle taken for each user occupying this vehicle (including pedestrians who are attached to the vehicles which hit them) - Numerical code |  |  |  |  |  |  |
| NUM\_VEH | NO | Identifier of the vehicle taken for each user occupying this vehicle (including pedestrians who are attached to the vehicles which hit them) - Alphanumerical code | Object | 0% |  |  |  |  |
| PLACE | NO | Allows you to locate the seat occupied in the vehicle by the user at the time of the accident. | FLOAT64 | 5,74% |  | 1 | 9 |  |
| CATU | NO | User category | int64 | 0% |  | 1 | 4 |  |
| GRAV | YES | Severity of user injury, injured users are classified into three categories (injured, hospitalized, killed) plus the uninjured | int64 | 0% |  | 1 | 4 |  |
| SEXE | NO | User gender | int64 | 0% |  | 1 | 2 |  |
| AN\_NAIS | NO | Year of birth of the user | FLOAT64 | 0,11% |  | 1900 | 2018 |  |
| TRAJET | NO | Reason for travel at time of accident | FLOAT64 | 0,02% |  | 0 | 9 |  |
| SECU1 | NO | The character information indicates the presence and use of safety equipment | FLOAT64 | 2,64% |  | -1 | 9 |  |
| SECU2 | NO | The character information indicates the presence and use of safety equipment | FLOAT64 |  |  | -1 | 9 |  |
| SECU3 | NO | The character information indicates the presence and use of safety equipment | FLOAT64 |  |  | -1 | 9 |  |
| LOCP | NO | Location of the pedestrian | FLOAT64 | 2,63% |  | -1 | 8 |  |
| ACTP | NO | Pedestrian action | FLOAT64 | 2,64% |  | -1 | 9 |  |
| ETATP | NO | This variable makes it possible to specify whether the injured pedestrian was alone or not: | FLOAT64 | 2,64% |  | -1 | 3 |  |

Some issues were identified in the data:

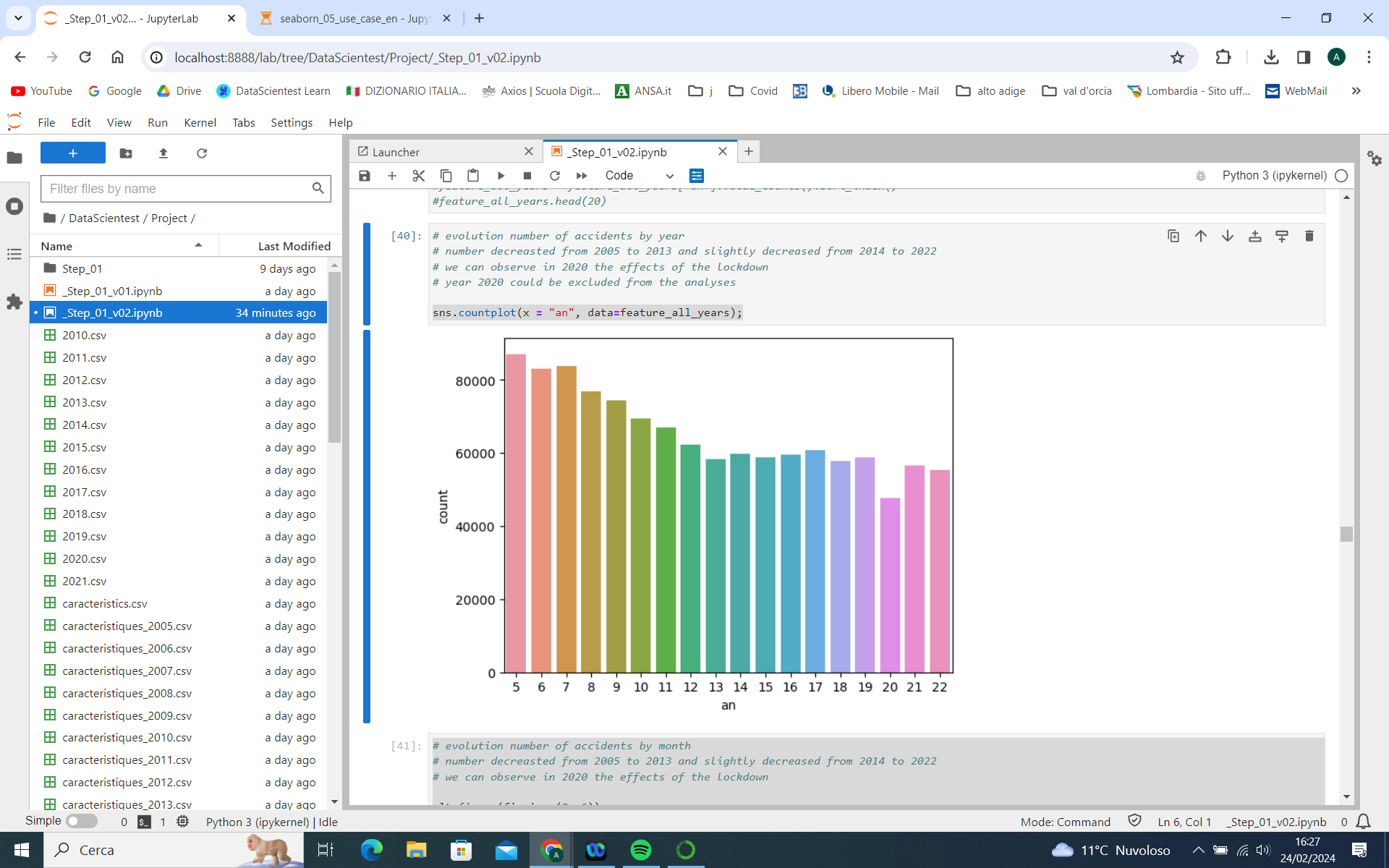
* In Characteristics files:
  + Hour/minute is available in format ‘hhmm’ (integer) or in ‘hh:mm’ (object)
  + com, dep with different formats (numerical, object). They also contain not correct values / errors
  + latitude and longitude have different data types. They have not been modified or manipulated at this stage
  + in some datasets ‘dep’ column contains values timed by 10 for inland departments. It has been necessary to normalize them dividing by 10 for values lower or equal to 950
* In Users files:
  + There are 29407 values for PLACE that could potentially be incorrect (double occupancy of one seat in a car)
* In Places files:
  + VOIE has different data types, some int and some float
  + There seems to be some typos or errors in the NBV values, e.g. for 561 missing values a space is entered before the -1 value, another value is #ERREUR
  + PR and PR1 have different data types, some int and some float. There are also some errors in the data, e.g., should be numerical value but some have parentheses
  + LARTPC has different data types, some int and some float. There are also some errors in the data, e.g., should be numerical value but some have parentheses

## Data Visualization

Further analysis has been done on the data, and some interesting patterns and trends have been observed:

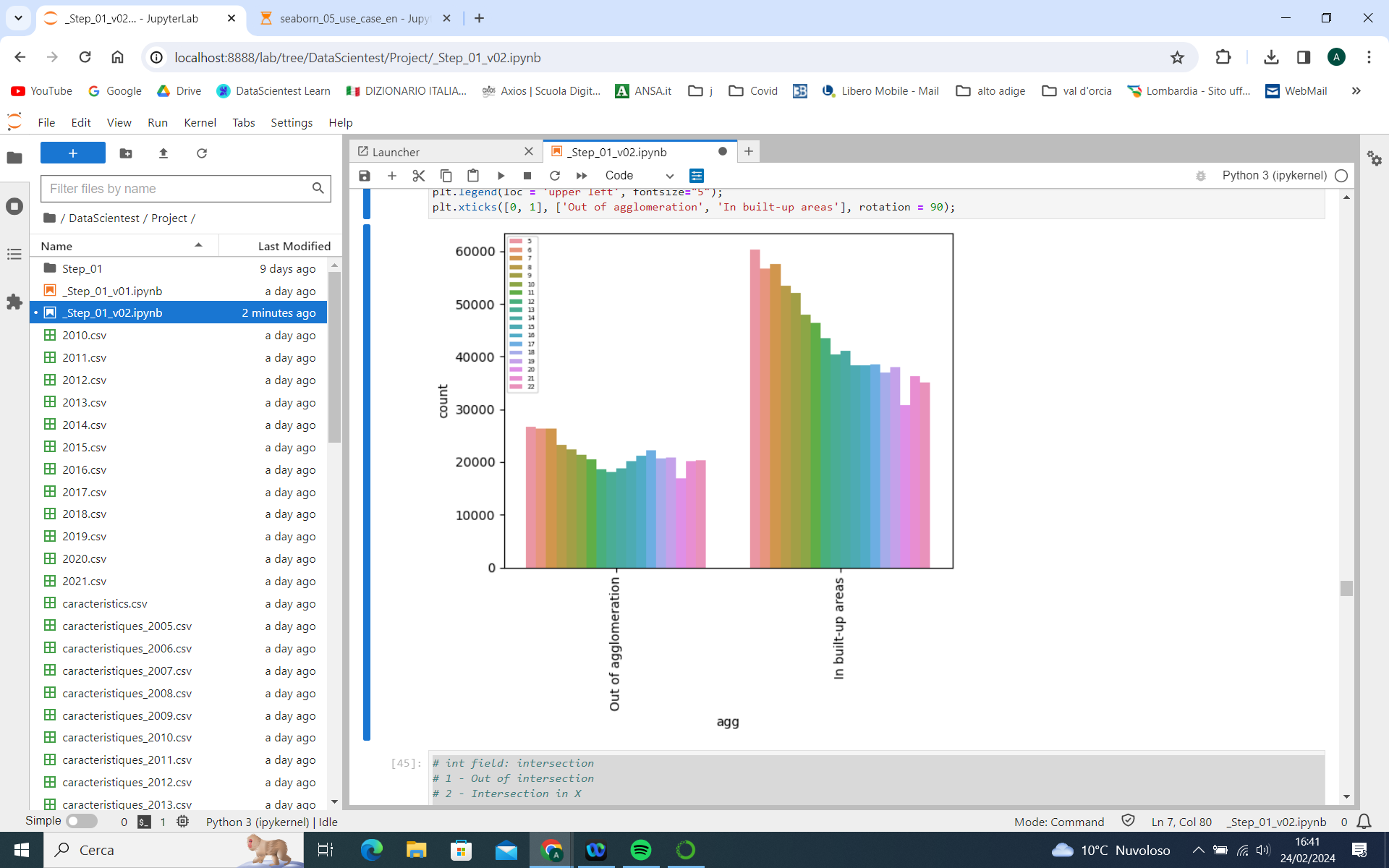
* Distribution of number of accidents per year

We observe a decrease of accidents in period between year 2005 and 2012. We can also see the effect of lockdown during the COVID pandemic in 2020 with a considerably lower number of accidents that can maybe be attributed to the reduction of car usage during that period:

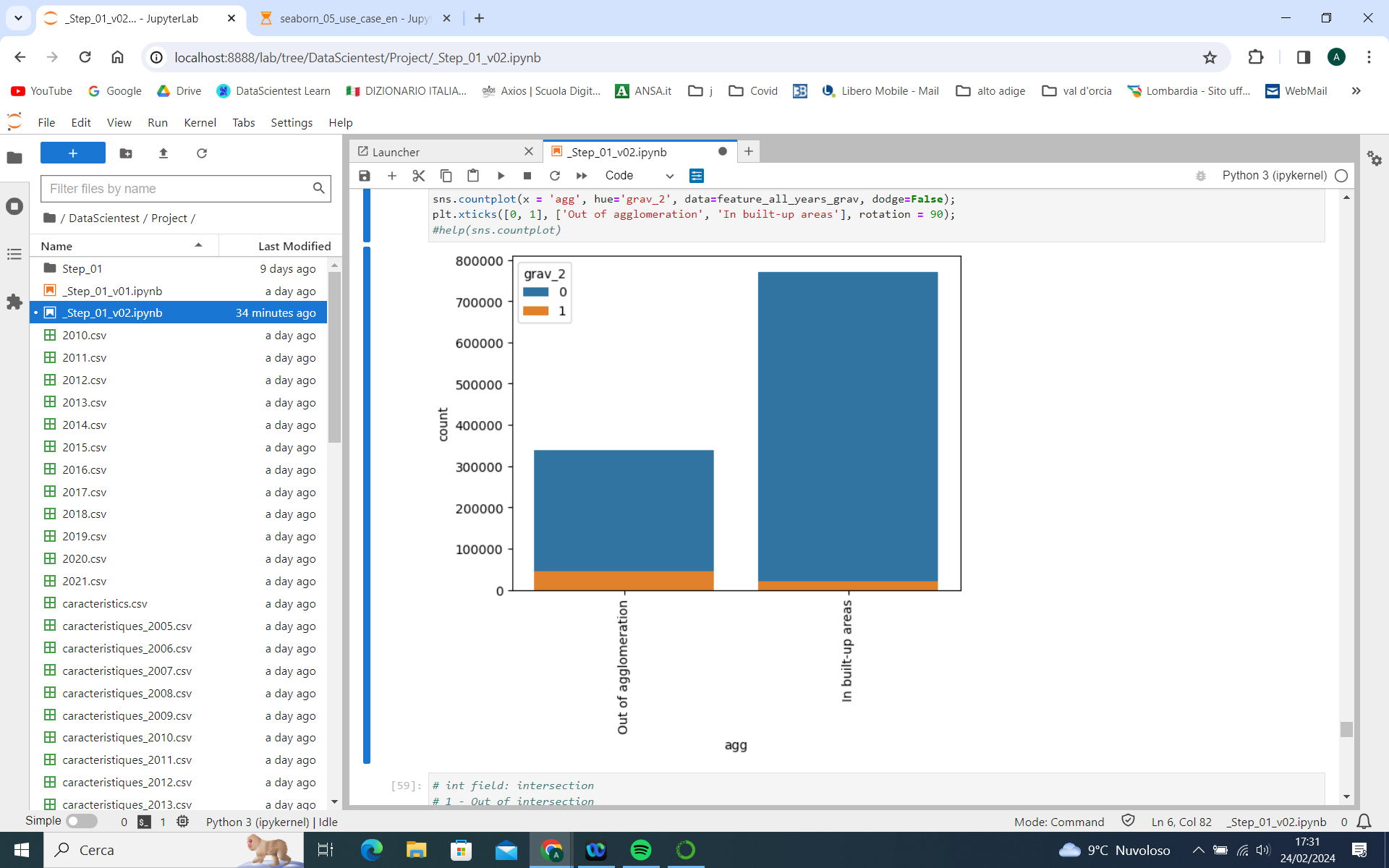


* Distribution of number and severity of accidents per location

The decrease in the number of accidents in period 2005-2012 looks to be especially observed in built-up areas (in comparison to areas out of agglomeration)



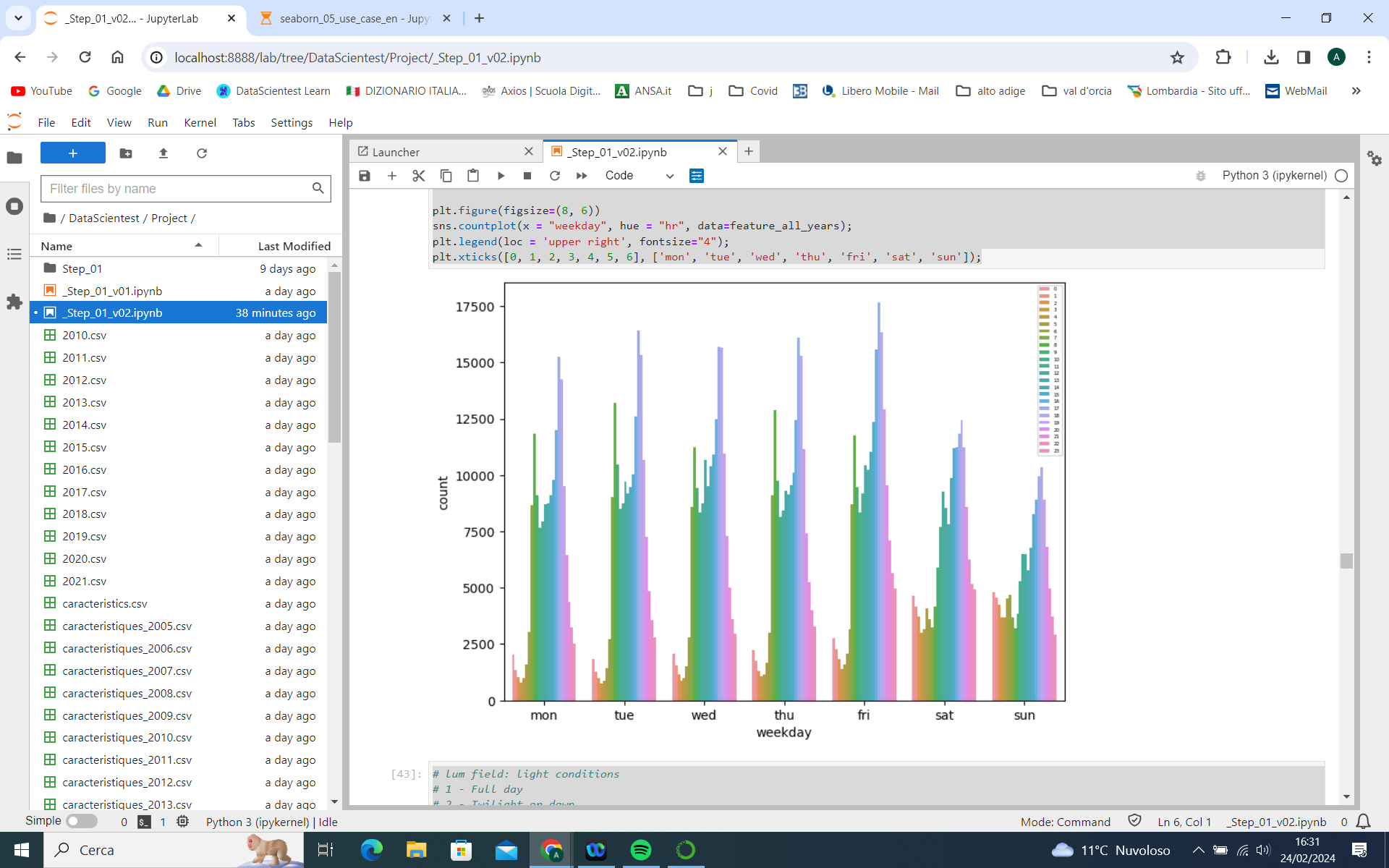
However, accidents in areas out of agglomeration lead to a higher incidences of death in comparison to built-up areas:



For this reason the decrease in the number of accidents observed from 2005 to 2012 might not necessarily lead to a proportional decrease of deaths

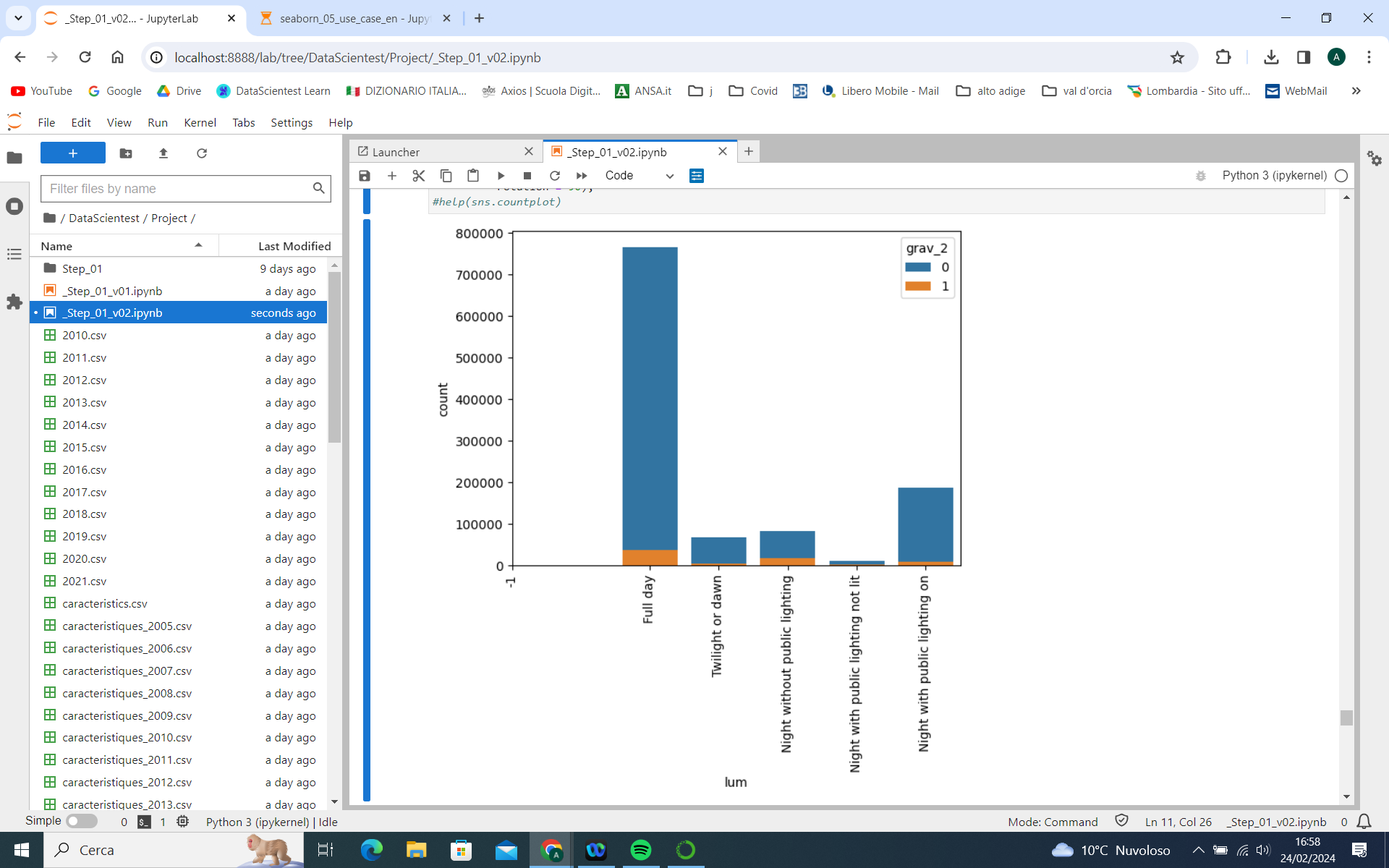
* Distribution of number of accidents per day and hour

We observe that the accidents are concentrated in peak hours of the day, especially in the evening. A higher number of accidents during the night is observed on Saturday and Sunday:



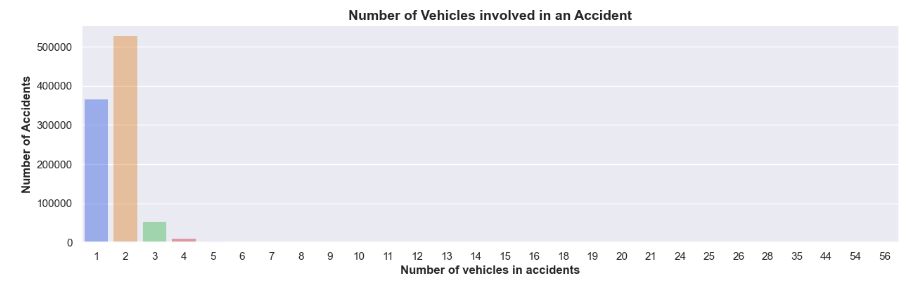
* Distribution of number of accidents and their severity by light condition:

We observe that most accidents occur during the day. However it is important to highlight a higher incidence of deaths during the night without public lighting:



* Distribution of number of accidents by number of vehicles involved:

We observe that the overwhelming majority of accidents involve 2 or 1 vehicles only:



* Number of accidents in every level of severity and by the number of Main maneuver before the accident.

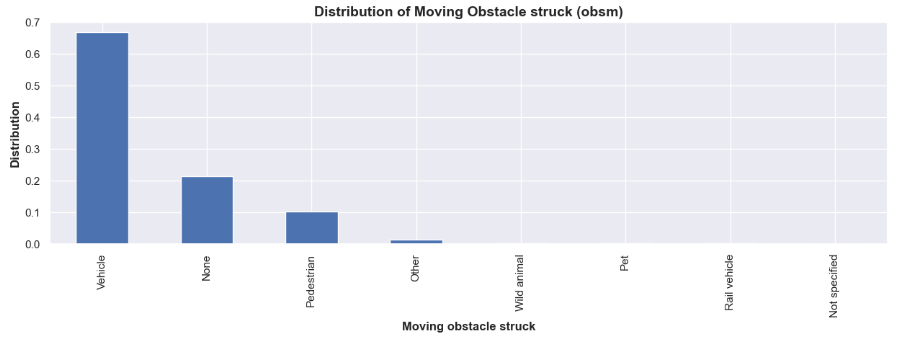
The characteristics of the variable main maneuver have been grouped into 5 categories. The highest level of severity accounts for most of the accidents without change of direction.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

* Distribution of number of accidents by type of moving obstacle hit:

We observe that the majority of accidents involve a vehicle hitting another moving vehicle which is in line with what we saw in the previous graph that the highest share of accidents involve 2 vehicles



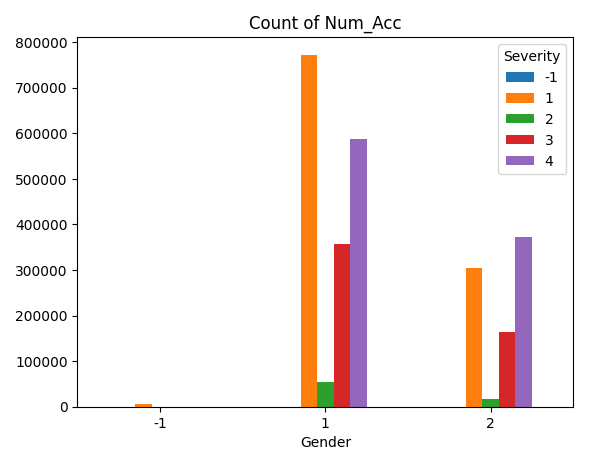
* Mortality rate based security equipment used:

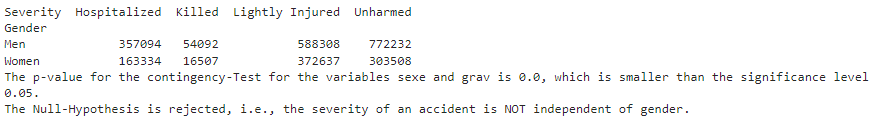
We observe that mortality is 10 times higher if the person is not wearing the seatbelt (18% in comparison to 1.8%). In the x-axis, 11 means wearing a belt and 12 means not wearing a belt:



* Distribution of number of accidents and their severity by gender:

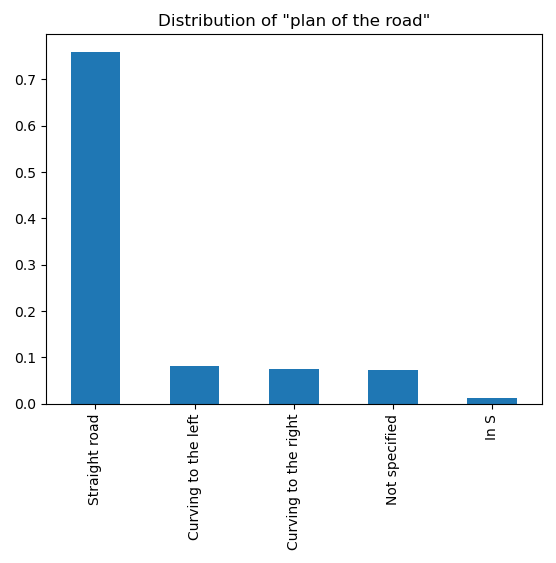
We observe a higher number of accidents involving male users (1 on x-axis) in comparison to female users (2 on x-axis). To evaluate the relationship between our target variable Severity and Gender, a chi square contingency test was then performed for gender and severity. The test resulted in a p-value of 0.0, which means that we can’t conclude that the severity of an accident is NOT independent from the gender.





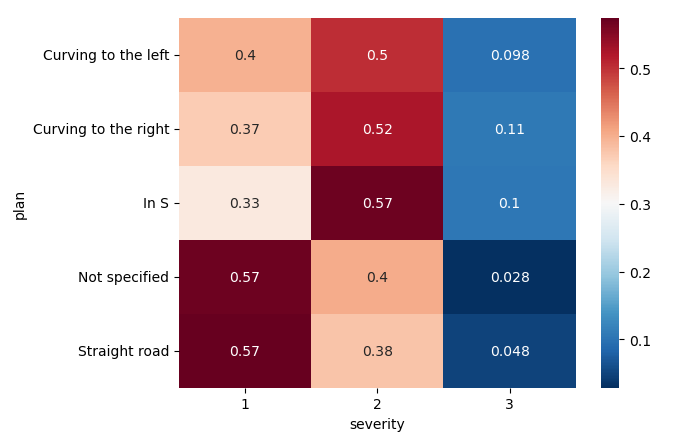
* Distribution of number of accidents and their severity by plan of the road:

We observe that the majority of accidents happen in roads with straight plan / layout (~75%) in comparison to roads curving to the left or right or serpent roads (in S shape):

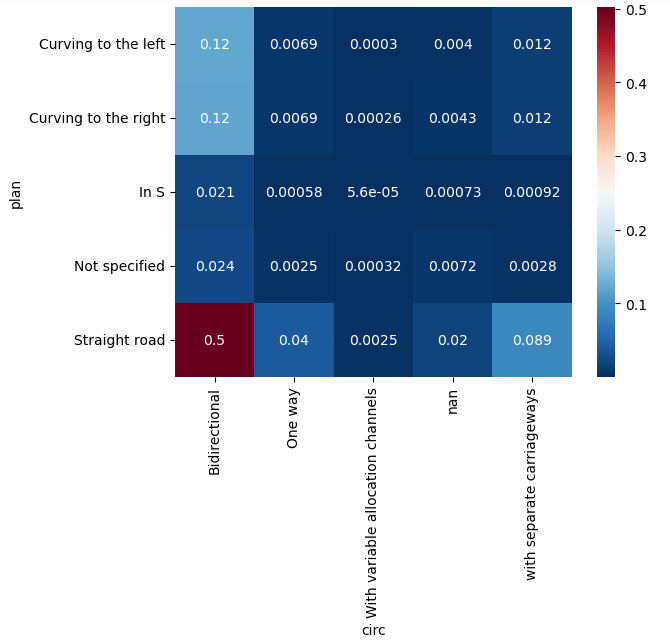


However, when we look at the severity of those accidents, we notice that accidents in curving or serpent roads are more likely to lead to death (severity 3 in the graph below). The chances that an accident on curving and serpent roads will lead to death if it happens in curving and serpent roads (9.8%, 10%, and 11%) is more than double the chances if it happens in straight roads (4.8%).

We can also see differences in injuries (severity 1 in graph below for few light injuries, 2 for strong injuries)



However, if we only want to look at the distribution of accidents that did lead to at least one death (i.e. not relative to the total number of accidents), we notice that almost 75% of them happen in bidirectional road, and more interestingly, we notice that half of all accidents that lead to at least one death did indeed happen on straight roads that are bidirectional:



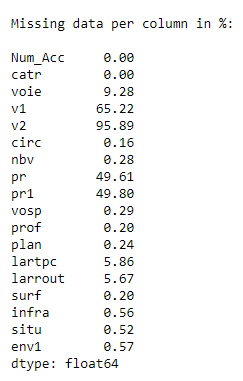
## Feature Engineering

This section will describe the feature engineering and the preprocessing steps performed on the data contained in all files of all 4 file types. We will describe the variables that we deem most relevant for achieving our project objectives and how we picked them. We will further include all preprocessing, formatting, and cleaning steps performed on the data in order to get to the final set of features we want to use for modelling.

### Locations file:

After examining missing values, we decided to drop columns with very high percentage of missing values. Namely columns “v1” and “v2”.

Columns “pr” and “pr1” contain in addition to missing values a lot of 0.0 and 0 values, all together add up to > 65% for “pr” and >70% for “pr1”. We therefore decided to drop them.

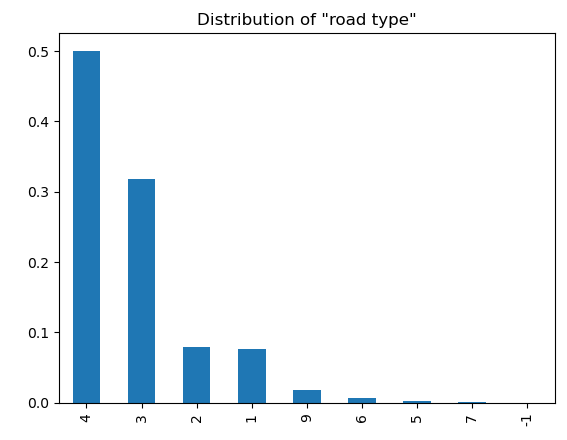


Furthermore, half the values in column “voie” are zeros, in addition to a further 9.2% missing values and other erroneous values. We therefore decided to drop this column as well.

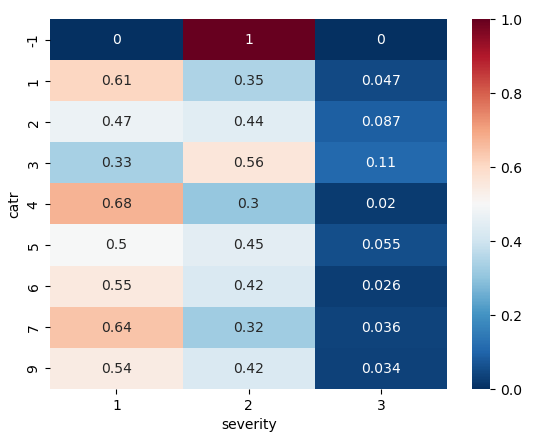
All remaining missing values are then replaced by a “-1”. Zeros “0” or “0.0” in most of the remaining columns indicate missing values as well and are therefore also replaced by a “-1”. In a small number of columns zeros might have an actual meaning and are not replaced. Those columns are:

* Infra: zero value can indicate that the road is none of the listed categories for infrastructure / development
* Vosp: zero values can indicate that the road does not have any of the listed reserved lane types
* Lartpc: zero values can indicate that the road does not have an island. So the width of the island might have been listed here as zero
* Larrout: zero values can indicate that the accident happened somewhere other than roads / driveway so the width of the driveway might have been listed here as zero.
* Column “catr”:

Majority of accidents happen in road types 4 (municipal road) and 3 (departmental road), followed by 2 (national road) and 1 (highway):

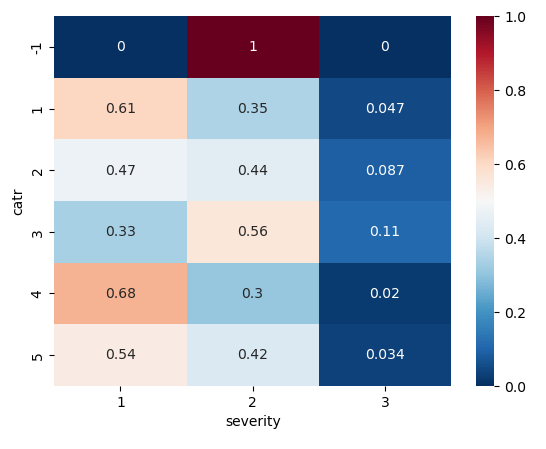


We reduce the number of categories in this column by grouping some of them with smallest sample size together based on similarity in nature and their similar correlation with target variable “accident severity”.



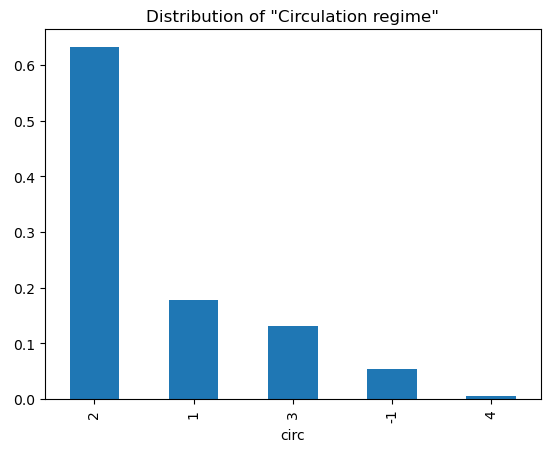
After examining the distribution of the different values with our target variable, the top 4 categories are kept as they are 4 for municipal road, 3 for departmental road, 2 for national road, and 1 for highway. The remaining categories 5, 6, 7, and 9 are grouped into 5 for others.

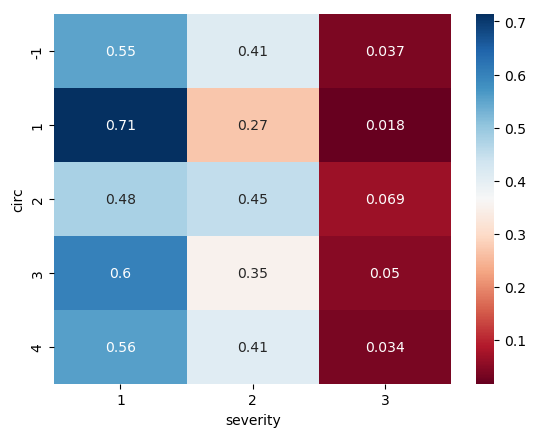
We check again the distribution of the different values with our target variable:



* Column “circ”:

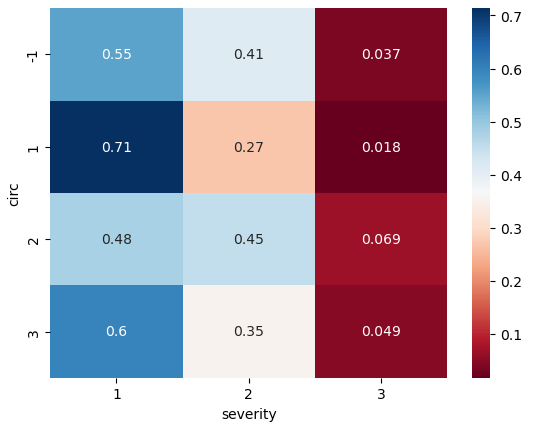
Majority of accidents happen in bidirectional roads (2 in the graph below), followed by one way roads (1) and then roads with separate carriageways (3) consecutively:





After examining the distribution of the different values with our target variable, the top 2 categories are kept as they are 2 for bidirectional roads and 1 for one way roads. The two remaining categories 3 (roads with separate carriageways) and 4 (roads with variable allocation channels) are grouped into 3 for others.

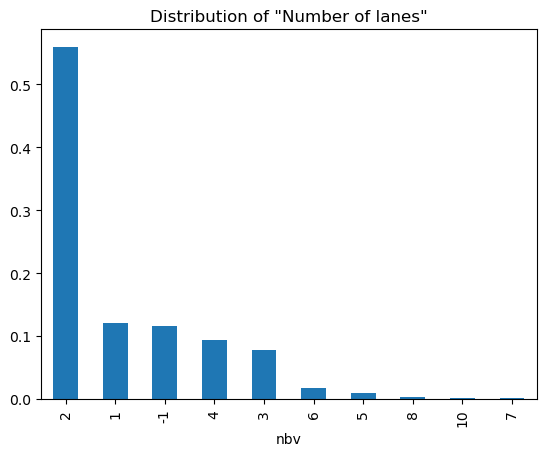
We check again the distribution of the different values with our target variable:

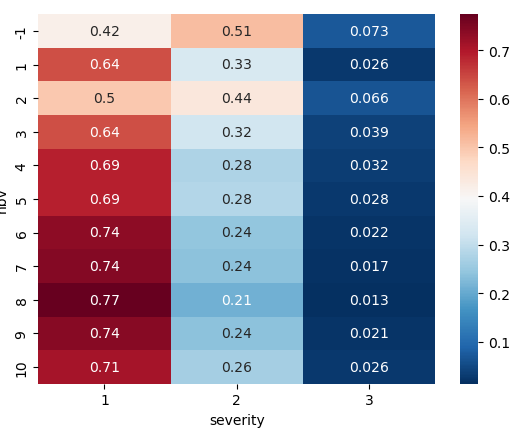


* Column “nbv”:

This column includes the number of lanes in the road in which the accident happened. It seems to include some erroneous values indicating an unreasonably high number of lanes. The values are then capped at 10, which means that any value higher than 10 is set to 10 instead.

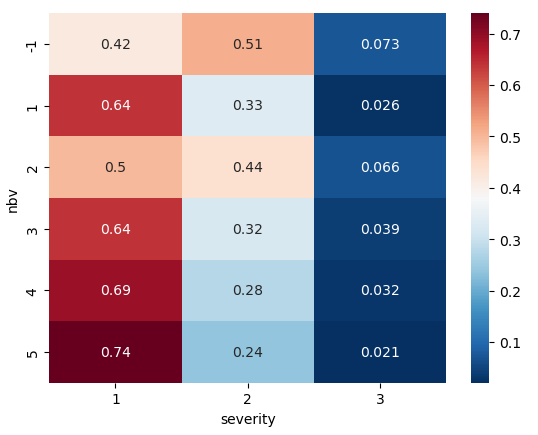
More than half of accidents happen in roads with 2 lanes. Followed byroads with one lane and then 4 land then 3 lanes consecutively:





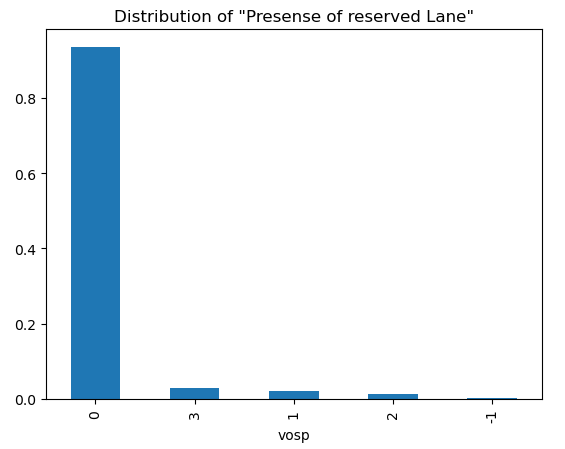
After examining the distribution of the different values with our target variable, values for roads with 1 lane, 2 lanes, and 3 lanes are kept the same. Roads with 4 or 5 lanes are grouped into one category “4” and roads with anything higher than 5 lanes are grouped together and given the value “5”.

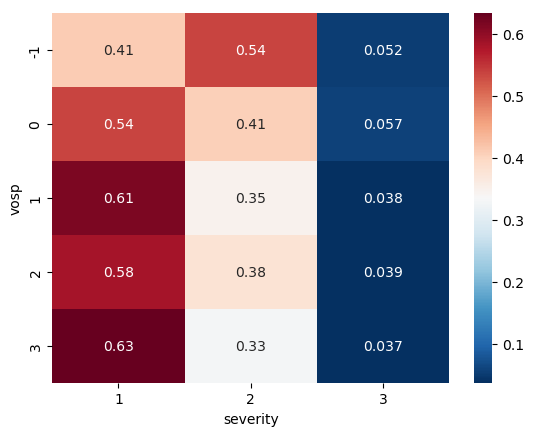
We check again the distribution of the different values with our target variable:



* Column “vosp”:

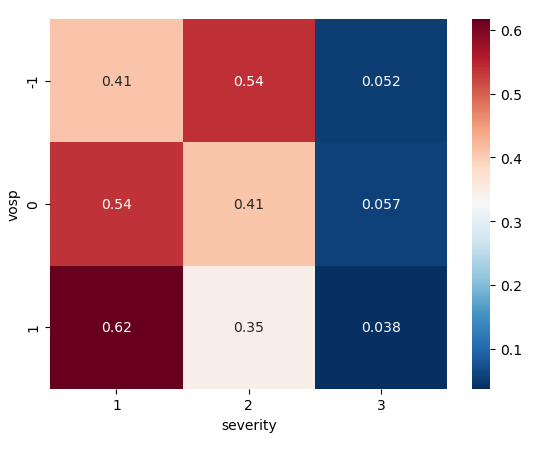
Almost 90% of accidents happen in roads that do not have a reserved lane (0 in the graph below):





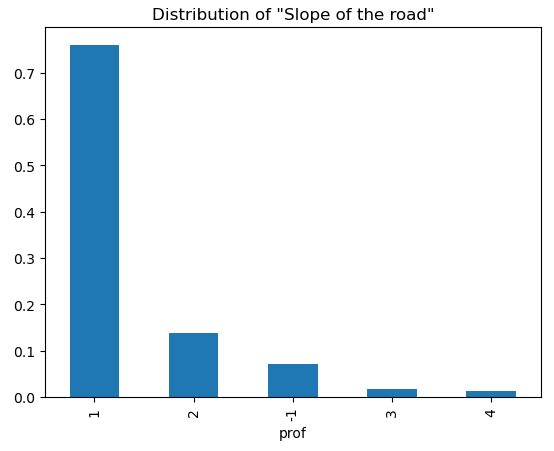
After examining the distribution of the different values with our target variable, this values are split into two values only; “0” for roads that dont have a reserved land or “1” for roads that have a reserved lane of any type.

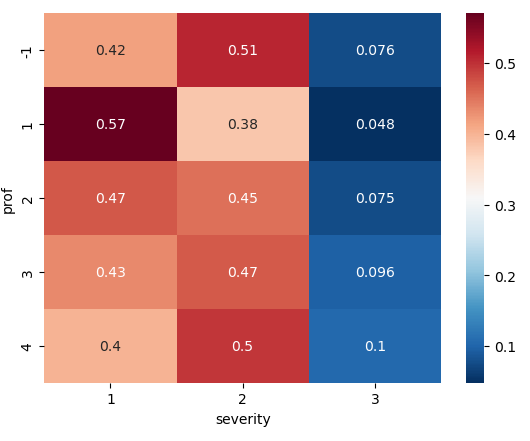
We check again the distribution of the different values with our target variable:



* Column “prof”:

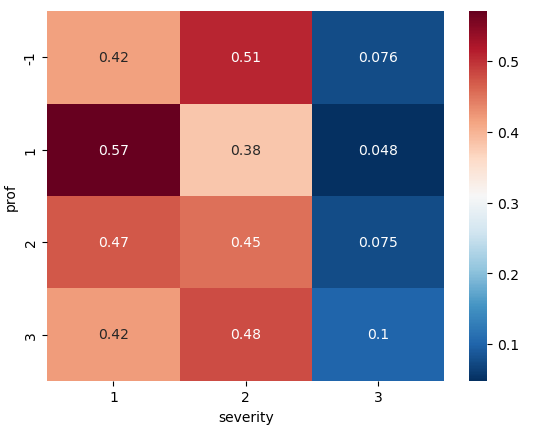
Majority of accidents happen in flat roads (1 in the graph below), followed by slope roads (2) and a small number of hill top roads (3) and hill bottom roads (4):





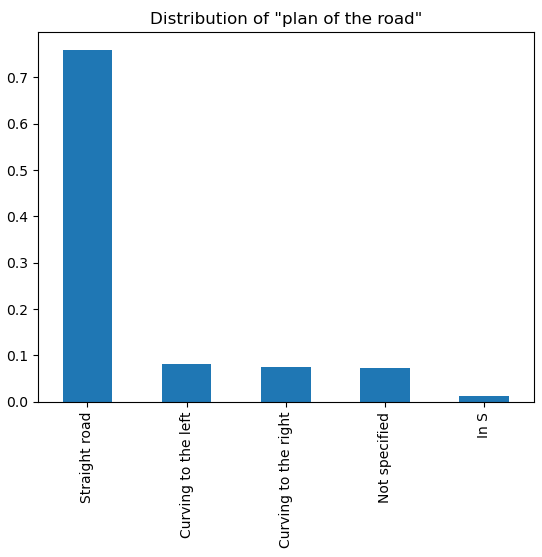
After examining the distribution of the different values with our target variable, the top two values are kept the same “1” for flat and “2” for slop. Hill top roads and hill bottom roads are grouped into one category “3”.

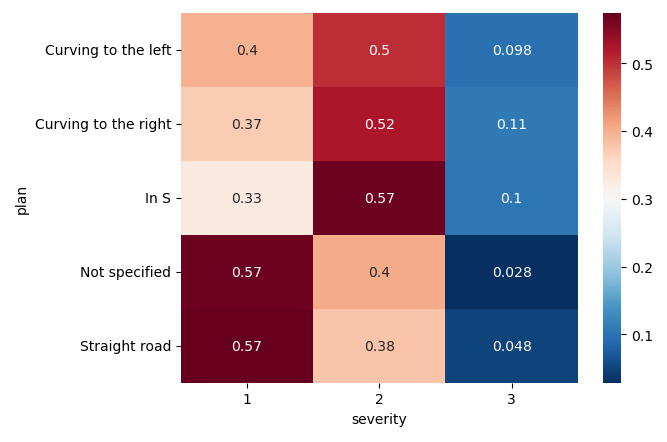
We check again the distribution of the different values with our target variable:



* Column “plan”:

Majority of accidents happen in straight roads, other categories each has a share less than 10% of accidents.





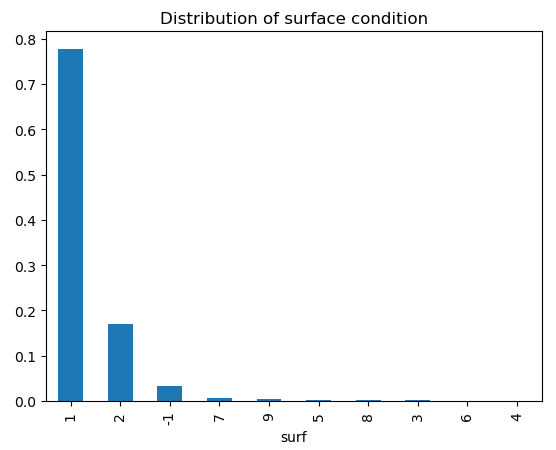
After examining the distribution of the different values with our target variable, the top value straight roads is kept the same “1”. All other values are grouped into one value “curvature” or “2”.

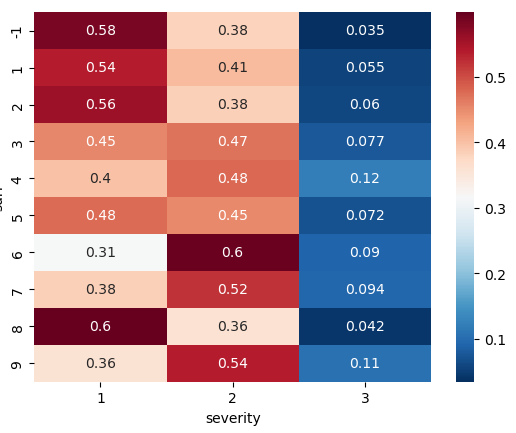
We check again the distribution of the different values with our target variable:



* Column “surf”:

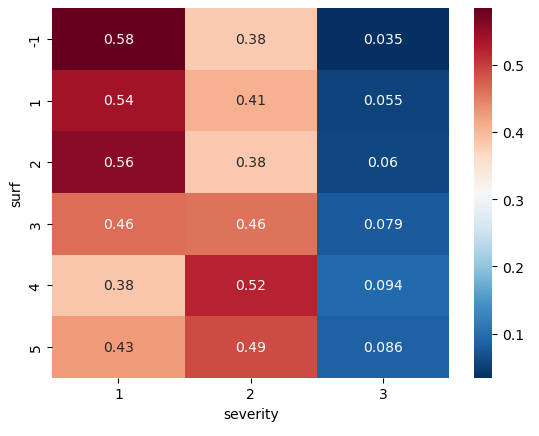
Majority of accidents happen in dry normal roads (1 in the graph below), followed by wet roads (2). All other categories account for a total of less than 10% of accidents.





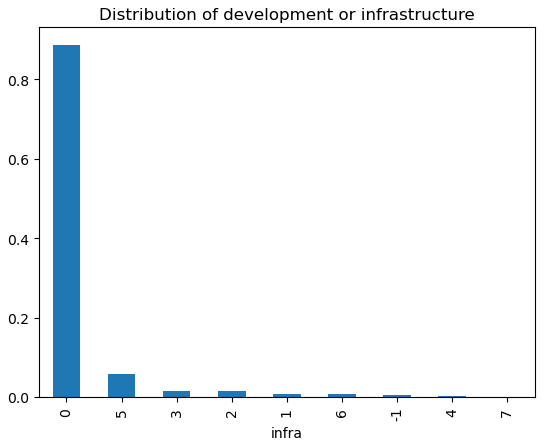
After examining the distribution of the different values with our target variable, the top 2 values for normal dry roads (1) and wet roads (2) stay the same. Roads with “puddles”, flooded roads, and snowy roads are grouped into one category (3). Icy roads are kept the same (4). All other values are grouped into one value “others” or “5”.

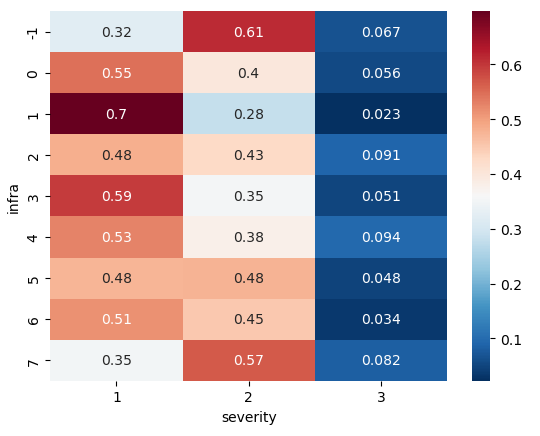
We check again the distribution of the different values with our target variable:



* Column “infra”:

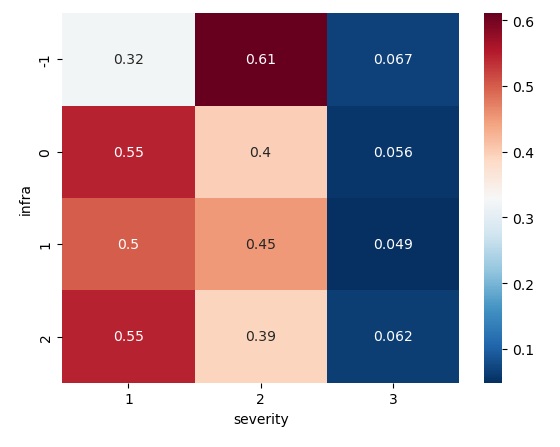
Majority of accidents ~85% happen in roads that are none of the listed development or infrastructure types (0 in the graph below). Followed by “intersections” (5) normal roads (1 in the graph below), followed by wet roads (2). All other categories account for a total of less than 10% of accidents.





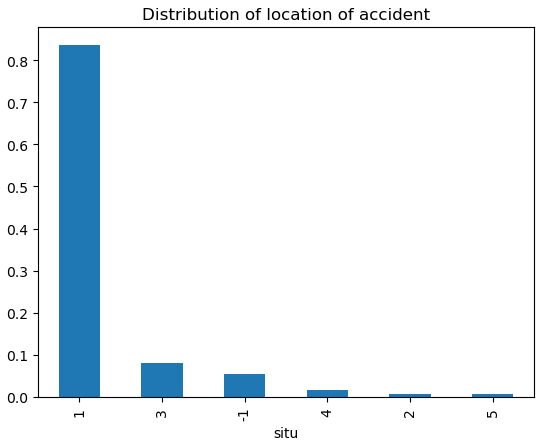
After examining the distribution of the different values with our target variable, the top value 0 stays the same. Interchange or connection ramps (3) and intersections (5) are grouped into one category “Intersections/Interchanges” (1) . All other values are grouped into one value “others” or (2).

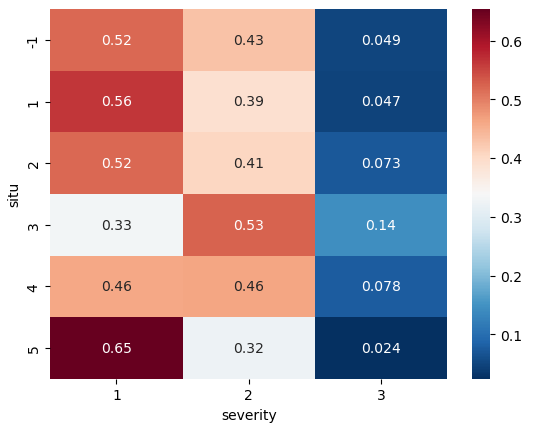
We check again the distribution of the different values with our target variable:



* Column “situ”:

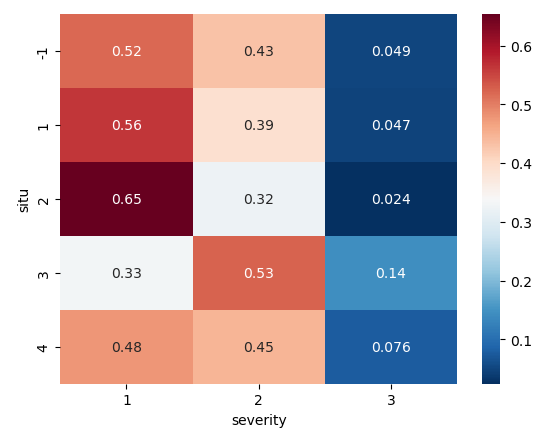
Majority of accidents ~85% happen on the driveways (1 in the graph below). Followed by “roadside / verge” (3). All remaining categories account for a total of less than 10% of accidents.





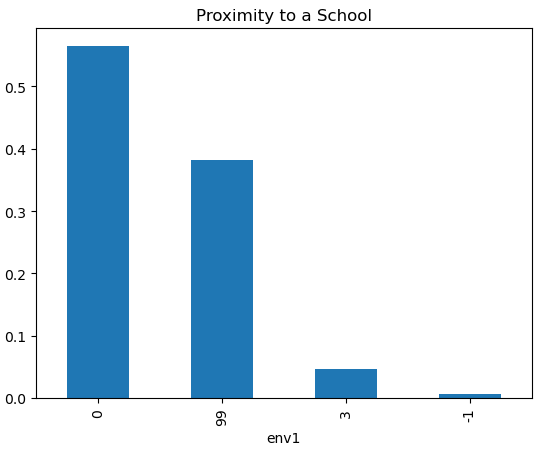
After examining the distribution of the different values with our target variable, the top 2 value “on driveway” (1) and “roadside / verge” (3) stay the same. Category “On cycle path” is switched from 5 to (2). All other values are grouped into one value “others” or (4).

We check again the distribution of the different values with our target variable:



* Column “env1”:

This column indicates the proximity of the accident location to a school.

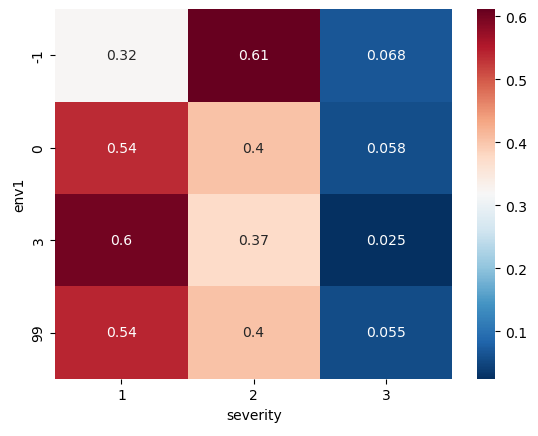


The meaning of the values in this column is not provided. Some assumptions were made:

Value 0 maybe indicates that there is no school nearby

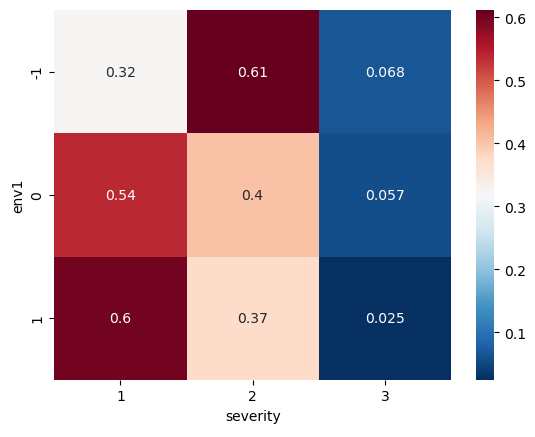
Value 3 maybe indicates that there is a school nearby, since it's different to value 0

Value 99: almost identical correlation with target variable when compared with value 0, maybe they mean the same and value 99 indicates that there is a no school nearby



After examining the distribution of the different values with our target variable, the value “1” is given instead of value “3” indicating that there is a school nearby. Value “0” is given to the two other values “0” and “99” indicating that there is no school nearby.

We check again the distribution of the different values with our target variable:



* Column “larrout”:

This column gives the width of the road / driveway where the accident happened. Some issues with the data were identified:

* Some roadway width is a negative value which doesn’t make sense, there were all made positive
* Some roadway width is entered in cm and some in meters, all values were transformed to cm
* Some values were below 250 cm, which is close to minimum requirements for road width. They will all replaced with 250

We then checked for outliers in this column:

We used the 0.9999 and 0.0001 quantiles to check for potential outliers. We got the value 250 for the lower quantile and the value 2450 for the upper quantile.

We then checked the outliers based on those quantiles, we get the following potential outliers (value and count of each):

|  |  |
| --- | --- |
| Larrout | Count |
| 2460 | 25 |
| 2480 | 14 |
| 2470 | 12 |
| 2490 | 5 |

We notice that there are no outliers lower than the 0.0001 quantile. They all equal the value 250. This can be explained by the fact that this lower quantile is close to the minimum requirement for road width.

We also notice that the potential outliers higher than the 0.9999 quantile are only slightly higher than this value. We therefore don't need to remove those values from our dataframe.

* Column “lartpc”:

This column gives the width of the island on the road where the accident happened. Some issues with the data were identified:

* Some island width is entered in cm, some in dm or meters,
* We will therefore transform all of them to cms:
* Some island width is entered in cm and some in dm or meters, all values were transformed to cm
* We will assume that the minimum width of an island is 15 cms. All values lower than 15 cm will be replaced with 15

We then checked for outliers in this column:

We used the 0.9999 and 0.0001 quantiles to check for potential outliers. We got the value 15 for the lower quantile and the value 900 for the upper quantile.

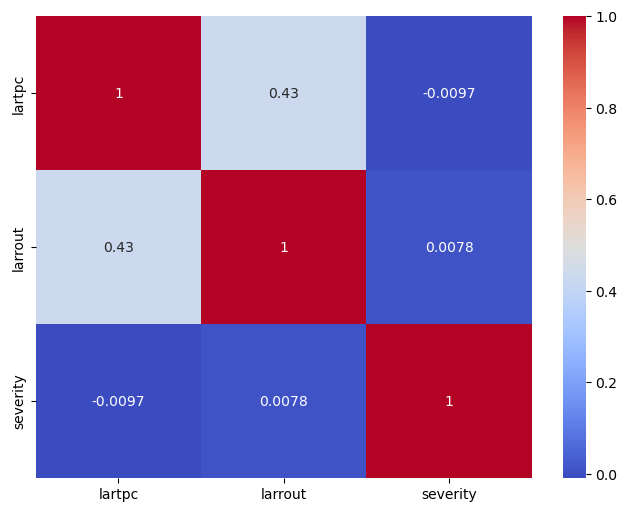
We then checked the outliers based on those quantiles, we get the following potential outliers (value and count of each):

|  |  |
| --- | --- |
| Larrout | Count |
| 907 | 2 |
| 905 | 2 |
| 910 | 1 |
| 960 | 1 |
| 950 | 1 |
| 915 | 1 |
| 906 | 1 |
| 904 | 1 |
| 908 | 1 |

We notice that there are no outliers lower than the 0.0001 quantile. They all equal the value 15. This can be explained by the fact that this lower quantile is the minimum possible value we set for island width

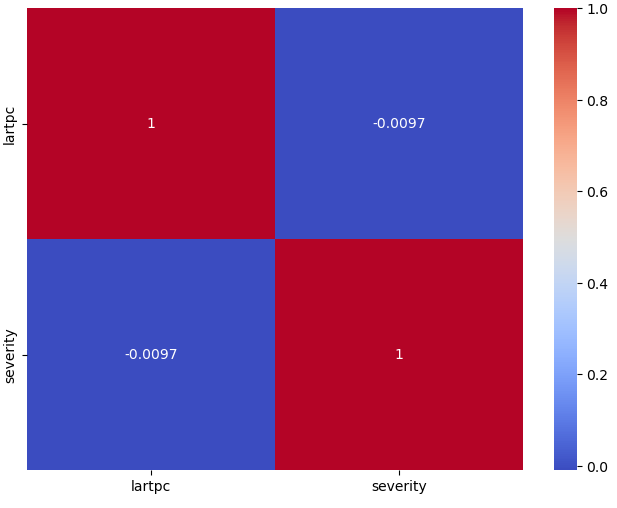
We also notice that the potential outliers higher than the 0.9999 quantile are only slightly higher than this value. We therefore don't need to remove those values from our dataframe.

We then evaluate the correlation of the two numeric variables “larrout” and “lartpc” with our target variable:



We notice that features “lartpc” and “larrout” have weak correlation with the target variable severity.

Because of this weak correlation and the fact that there is a large percentage of missing values, we can decide to drop off “larrout” going forward for the modelling steps. As for “lartpc” we can try to transform it into a binary categorical variable, “street has island” or “street without island” to see whether we can improve the correlation coefficient with the target variable.



After transformation, we see no improvement in the correlation coefficient to target variable. We can decide to drop off “lartpc” too going forward for the modelling steps.

* Feature selection:

Chi-square test: for our categorical variables we performed Chi-square test of independence of variables in a contingency table:

Test is done on one side using the target variable “Severity”

On the other side, we used each of the following columns, one at a time: “catr”, “circ”, “nbv”, “vosp”, “prof”, “plan”, “surf”, “infra”, “situ”, “env1”

Test results were the same for all columns:

Assumption(H0): The two columns are NOT related to each other

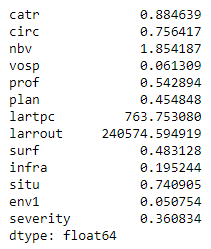
The P-Value of this Chisq Test is: 0.0

Result: P-Value < 0.05 then only we cannot accept the assumption (HO)

the tested variable is therefore correlated with our target variable severity

Variance tests:

We then checked the variance of our variables:



Correlation coefficient with target:

We then checked the correlation coefficient of our categorical variables with our target variable “severity”.

The bottom 5 variables with the lowest correlation coefficient to the target variable are:

* prof (0.0046)
* infra (0.005)
* surf (0.028)
* env1 (-0.041)
* vosp (-0.041)



Sample model and feature importance:

We ran a sample random forest model and we check the feature importance.

The 5 variables with the lowest feature importance based on the sample model we ran are:

* env1 (0.020231)
* vosp (0.021236)
* infra (0.038955)
* prof (0.057053)
* surf (0.061054)

Same 5 variables that had the lowest correlation coefficient to the target variable above.

Several attempts were made to change the grouping of categories for each of those 5 columns but no improvements could be achieved. Based on this and all previous tests, we decided to drop off the bottom 3 features in terms of feature importance going forward.

Columns “env1”, “vosp”, and “infra” we dropped.

Final list of features from “Locations” file:

* “catr”
* “circ”
* “nbv”
* “prof”
* “plan”
* “surf”
* “situ”

### Vehicles file:

We found out that the vehicles file didn’t contain a high percentage of missing values:

Missing data per column in %:  
  
Num\_Acc 0.00  
senc 0.02  
catv 0.00  
occutc 0.00  
obs 0.06  
obsm 0.05  
choc 0.02  
manv 0.03  
num\_veh 0.00  
dtype: float64

All remaining missing values are then replaced by a “-1”. Zeros “0” or “0.0” in most of the remaining columns indicate missing values as well.

* senc = Flow direction

Majority of accidents (>80%) happen with an unknown or not specified flow direction:

Unknown/Not specified 1359801  
PK or PR or ascending postal address number 172906  
PK or PR or decreasing postal address number 102832  
replaced missing value 272

Ein Bild, das Screenshot, Rechteck, Quadrat, Design enthält.

Automatisch generierte Beschreibung

This is already an indication for the variable not being relevant. But we will investigate it later.

* obs = Fixed obstacle hit

Analog to the variable senc, the majority of accidents (>80%) happen where the obstacle hit is unknown or not specified: We can assume that this variable is irrelevant but deeper analysis will follow later.

Unknown/Not specified 1421430  
Parked vehicle 35471  
Ditch, embankment, rock wall 27498  
Tree 23414  
Concrete slide 18712  
Metal slide 18225  
Building, wall, bridge pier 18018  
Post 17018  
Other fixed obstacle on the roadway 11974  
Curbside 9223  
Exiting the roadway without obstacles 9055  
Other fixed obstacle on sidewalk or shoulder 7743  
Urban furniture 5287  
Island, refuge, upper boundary 3901  
Vertical signaling support or emergency call station 3674  
Other slide 2247  
Parapet 1915  
replaced missing value 1006  
Name: obs\_label, dtype: int64

Ein Bild, das Screenshot, Rechteck, Display, Quadrat enthält.

Automatisch generierte Beschreibung

* obsm = Moving obstacle struck

After examining the distribution of the obsm variable, that vehicles are the most distributed moving obstacle struck in an accident. This category is followed by unknown values and Pedestrian,

Vehicle 1090958  
Unknown/Not specified 347714  
Pedestrian 167381  
Other 23174  
Wild animal 2760  
Pet 1529  
Rail vehicle 1517  
replaced missing value 778  
Name: obsm\_label, dtype: int64

Ein Bild, das Screenshot, Rechteck, Quadrat, Design enthält.

Automatisch generierte Beschreibung

* choc = Initial shock point

After examining the initial shock point of an vehicle during an accident it shows that the categories referring to the front part of the vehicle are the most common shock points.

Before 604020  
Front left 234482  
Front right 189967  
Back 153497  
Left side 119967  
Unknown/Not specified 109680  
Right side 100263  
Left rear 54703  
Right back 42960  
Multiple impacts (rollovers) 25875  
replaced missing value 397  
Name: choc\_label, dtype: int64

Ein Bild, das Screenshot, Grafiken, Design enthält.

Automatisch generierte Beschreibung

* manv = Main maneuver before the accident

Ein Bild, das Screenshot, Rechteck, Reihe, Design enthält.

Automatisch generierte Beschreibung

We reduce the number of categories in this column by grouping some of them together. This grouping is based on the similarity of maneuvers and their potential impact on traffic flow and accident likelihood.

* **No Change of Direction:**
  + 1: Without change of direction
  + 2: Same direction, same line
  + 3: Between 2 lines
  + 7: In the bus lane, in the same direction
  + 11: Changing lanes Left
  + 12: Changing lanes Right
  + 13: Deported Left
  + 14: Deported Right
  + 17: Exceeding Left
  + 18: Exceeding Right
* **Wrong Direction**
  + 4: In reverse
  + 5: Against the grain
  + 8: In the bus lane, in the opposite direction
* **Change of Direction:**
  + 6: Crossing the central reservation
  + 10: When making a U-turn on the road
  + 15: Turning point Left
  + 16: Turning point Right
  + 19: Crossing the road
* **Special Maneuvers:**
  + 9: By inserting yourself
  + 20: Parking maneuver
  + 21: Avoidance maneuver
  + 22: Door opening
  + 23: Stopped (excluding parking)
  + 24: Parked (with occupants)
  + 25: Driving on the sidewalk
  + 26: Other maneuvers
* **Uncategorized:**
  + -1: Replaced missing value
  + 0: Unknown/Not specified

The new grouping shows, that no change of directions is the most common value in an accident, followed by a change of direction.

Ein Bild, das Screenshot, Rechteck, Grafiken, Quadrat enthält.

Automatisch generierte Beschreibung

* Num\_Acc:

Identifier of the accident identical to that of the “CHARACTERISTICS ” file included for each of the vehicles described involved in the accident.

After examining this variable, we can see that in most of the accidents 2 vehicles were involved, whereas still in a considerable number of accidents only one car was involved.

Ein Bild, das Text, Diagramm, Reihe, Screenshot enthält.

Automatisch generierte Beschreibung

* catv = Vehicle category

Ein Bild, das Text, Screenshot, parallel enthält.

Automatisch generierte Beschreibung

Transformation of CATV column:

based on the provided categories for the variable **catv**, we create groups. This grouping is based on the similarity of vehicle types and their usage:

* **Motorcycles:**
  + 1: Bicycle
  + 2: Moped <50cm3
  + 30: Scooter < 50 cm3
  + 31: Motorcycle > 50 cm3 and <= 125 cm3
  + 32: Scooter > 50 cm3 and <= 125 cm3
  + 33: Motorcycle > 125 cm3
  + 34: Scooter > 125 cm3
  + 35: Light Quad <= 50 cm3 (Quadricycle with unbodied engine)
  + 36: Heavy quad > 50 cm3 (Quadricycle with non-bodied engine)
  + 41: 3WD <= 50 cm3
  + 42: 3WD > 50 cm3 <= 125 cm3
  + 43: 3WD > 125 cm3
* **Cars and Similar Vehicles:**
  + 3: Carriage (Quadricycle with body motor) (formerly “motor car or tricycle”)
  + 7: VL only
  + 8: Reference unused since 2006 (VL + caravan)
  + 9: Reference unused since 2006 (VL + trailer)
  + 10: LCV only 1.5T <= GVW <= 3.5T with or without trailer (formerly LCV only 1.5T <= GVW <= 3.5T)
  + 13: PL only 3.5T <PTCA <= 7.5T
  + 14: PL only > 7.5T
  + 15: PL > 3.5T + trailer
  + 16: Road tractor alone
  + 17: Road tractor + semi-trailer
  + 37: Bus
  + 38: Coach
* **Public Transport:**
  + 18: Reference unused since 2006 (public transport)
  + 19: Reference unused since 2006 (tram)
  + 39: Train
  + 40: Tram
* **Specialized Vehicles:**
  + 20: Special device
  + 21: Agricultural tractor
  + 50: Motorized EDP
  + 60: EDP without motor
  + 80: VAE
* **Others:**
  + 4: Reference unused since 2006 (registered scooter)
  + 5: Reference unused since 2006 (motorcycle)
  + 6: Reference unused since 2006 (sidecar)
  + 11: Reference unused since 2006 (UV (10) + caravan)
  + 12: Reference unused since 2006 (UV (10) + trailer)
  + 99: Other vehicle

The results depict the distribution of transportation modes: cars dominate at approximately 71.46%, followed by motorcycles at 25.77%, with specialized modes making up a smaller proportion at 2.50%. Public transport represents a minimal fraction, accounting for only about 0.27% of the total.

Ein Bild, das Text, Diagramm, Reihe, Rechteck enthält.

Automatisch generierte Beschreibung

* occutc: Number of occupants on public transport.

Ein Bild, das Text, Screenshot, Reihe enthält.

Automatisch generierte Beschreibung

Proportion Accident  
with public transportation: 0.68 %  
without public transportation: 99.32 %

These results suggest that 0.68% of accidents involve public transportation, while the vast majority, 99.32%, occur without any involvement of public transportation. This indicates that accidents primarily occur in contexts unrelated to public transport, highlighting the comparatively lower incidence of accidents involving public transportation systems.

* num\_veh

*Vehicle identifier taken for each user occupying this vehicle (including pedestrians who are attached to the vehicles which hit them) – Alphanumeric code*

Ein Bild, das Text, Diagramm, Reihe, Quittung enthält.

Automatisch generierte Beschreibung

Correlation coefficient with target:

We then checked the correlation coefficient of our categorical variables with our target variable “severity”.

Ein Bild, das Text, Screenshot, Zahl, parallel enthält.

Automatisch generierte Beschreibung

When comparing the correlation coefficients with the target variable "grav" (severity of accidents), we can observe:

* senc and grav: 0.0062 (very weak positive correlation)
* catv and grav: -0.18 (weak negative correlation)
* occutc and grav: -0.01(very weak negative correlation)
* obs and grav: 0.1 (weak positive correlation)
* obsm and grav: -0.099 (weak negative correlation)
* choc and grav: -0.043 (very weak negative correlation)
* manv and grav: -0.14 (weak negative correlation)

Given this context:

* The type of vehicle (catv) shows a negative correlation with the severity of accidents (grav). Certain types of vehicles might be associated with more severe accidents.
* The number of moving obstacles (obsm) has a weak negative correlation with the severity of accidents (grav). This implies that, on average, accidents involving moving obstacles might be slightly less severe.
* The maneuver of the driver (manv) still has a weak negative correlation with the severity of accidents (grav). Certain maneuvers might be associated with less severe accidents.
* obs and obsm: There is a strong negative correlation between observation points (obs) and observation of motorcycles (obsm) (-0.3).
* choc and manv: There is a moderate positive correlation between the point of impact (choc) and the maneuver of the driver (manv) (0.15).
* occutc and obs: There is a positive correlation between the number of people involved (occutc) and observations (obs) (0.0067).

The variables with a correlation to the target variable below 0.01 are:

['Num\_Acc', 'senc', 'occutc', 'obsm', 'choc']

Sample model and feature importance:

We ran a sample random forest model and we check the feature importance. The variables with a feature importance below 0.01 are:

'manv', 'occutc', 'senc'

Based on this and all previous tests, we decided to drop off least relevant according to feature importance < 0.1"

We also decided to drop OBS column, as > 86% of the values in this column are of category "Unknown or Unspecified" => It won't add value to the prediction

Conclusion for Irrelevant columns in Vehicles file:

**occutc**

* does not show any meaningful correlation or relationship with the target variable or other relevant variables (correlation with target variable: -0.01)
* variable has the same value for almost all observations in dataset, it provides no useful information for analysis (99.32 % of the data has value 0)

**senc**

* does not show any meaningful correlation or relationship with the target variable or other relevant variables (correlation with target variable: 0.0062)
* It might be considered relatively less important in predicting accident severity
* variable has the same value for almost all observations in dataset, it provides no useful information for analysis (>83 % of the data has value "unknown/not specified")

**manv**

* does not show any meaningful correlation or relationship with the target variable or other relevant variables (correlation with target variable: -0.0923)
* It might be considered relatively less important in predicting accident severity

**obs (fixed obstacle).**

* show weak correlation (at threshold) with severity of accidents. However, distribution of column shows > 86% of values belong to category "unknown"

**Additional Transformation:**

* Drop categories 3 & 4 of variable catv => 0,5% values belong to either 3 or 4 category.
* Drop categories -1, 4, 5 & 6 of variable obs => <0,5% values belong to those categories

Statistical Tests

we used the Chi-Square test of independence, followed by Cramer's V, which measures the level of dependence between variables.

To do this, we used the contingency table, which counts the cross-modalities between our two variables. From this table we can perform our test of independence.

The Chi-Square test enables us to test the dependence of two categorical variables. Two hypotheses are considered :

* H0 : There is no association between the two variables,;
* H1 : There is an association between the two variables,.

When the p-value of the statistical test is less than 0.05, we reject the null hypothesis (H0) at 95%, and our variables are considered dependent.

The Chi-Square test lets us know whether our two variables are significantly dependent. If they are, and we wish to quantify this dependence, we can then calculate it using Cramer's V, which takes as its argument the value of the Chi-Square test statistic.

The formula for calculating Cramer's V is V=sqrt(𝜒2/(𝑁×[𝑚𝑖𝑛(𝑎,𝑏)−1]))

where 𝜒2 corresponds to the test statistic, a to the number of categories in the first variable, b to the number of categories in the second variable and N to the number of dataset rows.

for the variable senc, the p-value is not below 0.05 => we can't reject our H0, suggesting that there is no significant association between the corresponding variable and grav. for all other variables, our p-value is below 0.05 => we reject the null hypothesis, suggesting that there is a significant association between the corresponding variable and grav.

interpretation the results based on the Cramer's V values and associated p-values:

* **catv (type of vehicle):**
  + **V\_Cramer:** 0.375509 (High)
  + **Chi2 p-value:** 0.0
* **Interpretation:** There is a strong and statistically significant association between the type of vehicle involved in an accident and the severity of accidents. The high Cramer's V value indicates a substantial effect size, suggesting that the type of vehicle has a significant impact on the severity of accidents.
* **choc:**
  + **V\_Cramer: 0.212 (Medium)**
  + **Chi2 p-value: 0.0**

**Interpretation:**The association between 'choc' and the target variable is statistically significant, similarly to the other variables.

* **Variable 'obsm':**
  + **V\_Cramer: 0.194 (Medium)**
  + **Chi2 p-value: 0.0**

**Interpretation:**The association between 'obsm' and the target variable is statistically significant, similarly to the other variables.

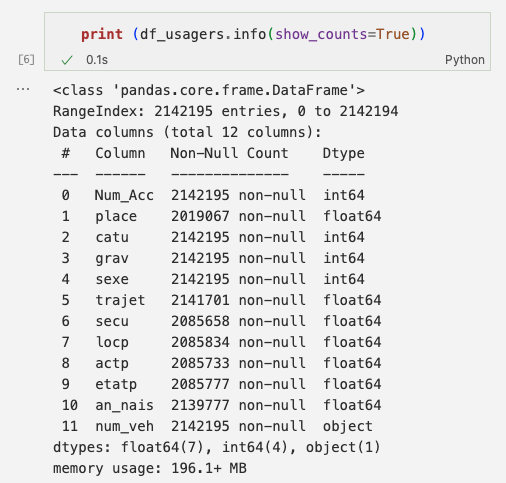
In summary, while all variables ('catv', 'choc', 'obsm') show statistically significant associations with the target variable, 'catv' demonstrates the strongest association, followed by 'choc' and 'obsm', with 'manv' showing the weakest association among the variables analyzed.

Final list of features from “Vehicles” file:

* 'Num\_Acc',
* 'catv',
* 'choc',
* ‘obsm’
* 'num\_veh'

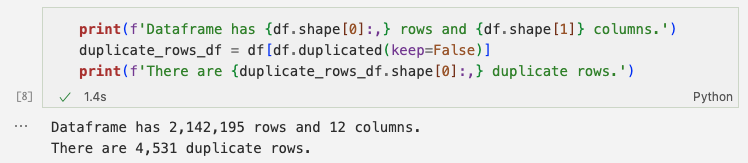
### Users file

#### Overview

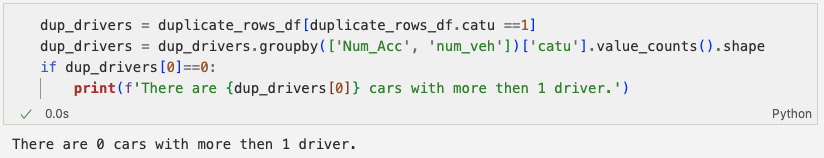
  
More than 2 Million rows in 12 columns. Some Columns have missing values.

#### Duplicates

#### Rows

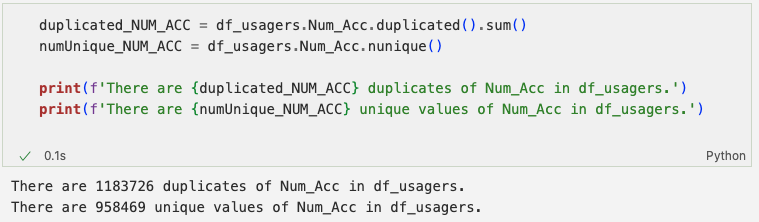


Duplicate rows are possible, only the driver seat of each car should not be occupied by more than 1 person. Check, if there is only 1 driver per vehicle:



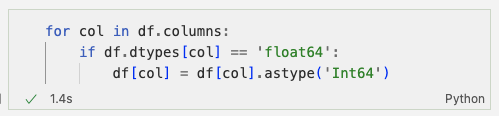
Result: Duplicate rows do not need to be removed.

#### Keys (Num\_Acc)



There are Accidents “Num\_Acc” with more than 1 record in df\_usagers: they will have to be merged to one row in order to join them with the other tables (caracteristiques, lieux, vehicles). This requires all variables that are not dropped, to be transformed into dummy-variables with reasonable splits, i.e., categories.

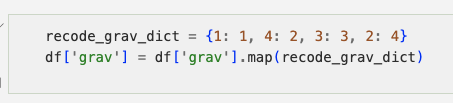
#### Transformation Type of Float-Columns to nullable Integer-Type “Int64”

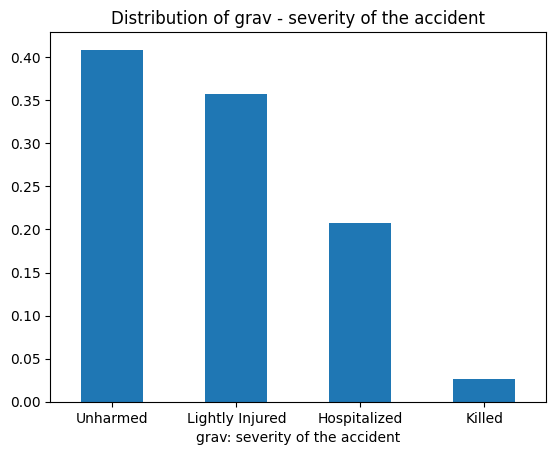


Note the capital “I” in Int64. Python’s int64-type doesn’t handle NaN-Values.

#### Column grav (severity of injury of the user)

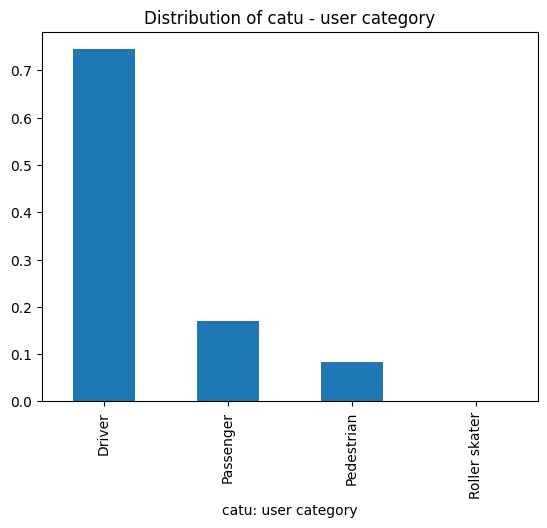
Recode the grav to reflect the severity of the accident in an ordinal variable. In the final dataframe (on row per Num\_Acc) those values will be summed up giving a measure for the severity of the accident, the mission target.





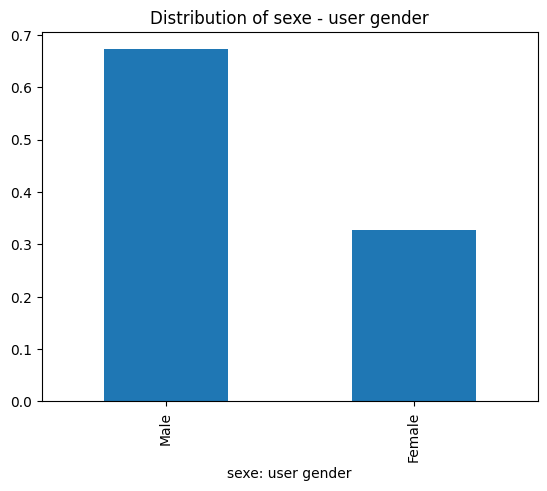
#### Column catu (user category)

Final\_DF: catu will be replaced by dummies, one for every catu\_category, counting the users per category. No cleaning needed. (We will later see, that this column is dropped due to strong correlation with place-variable)

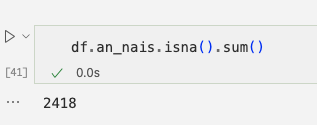


#### Column Sexe (Gender)

Final\_DF: sexe will be replaced by dummies, one for every sexe\_category, counting the users per category. No cleaning needed.

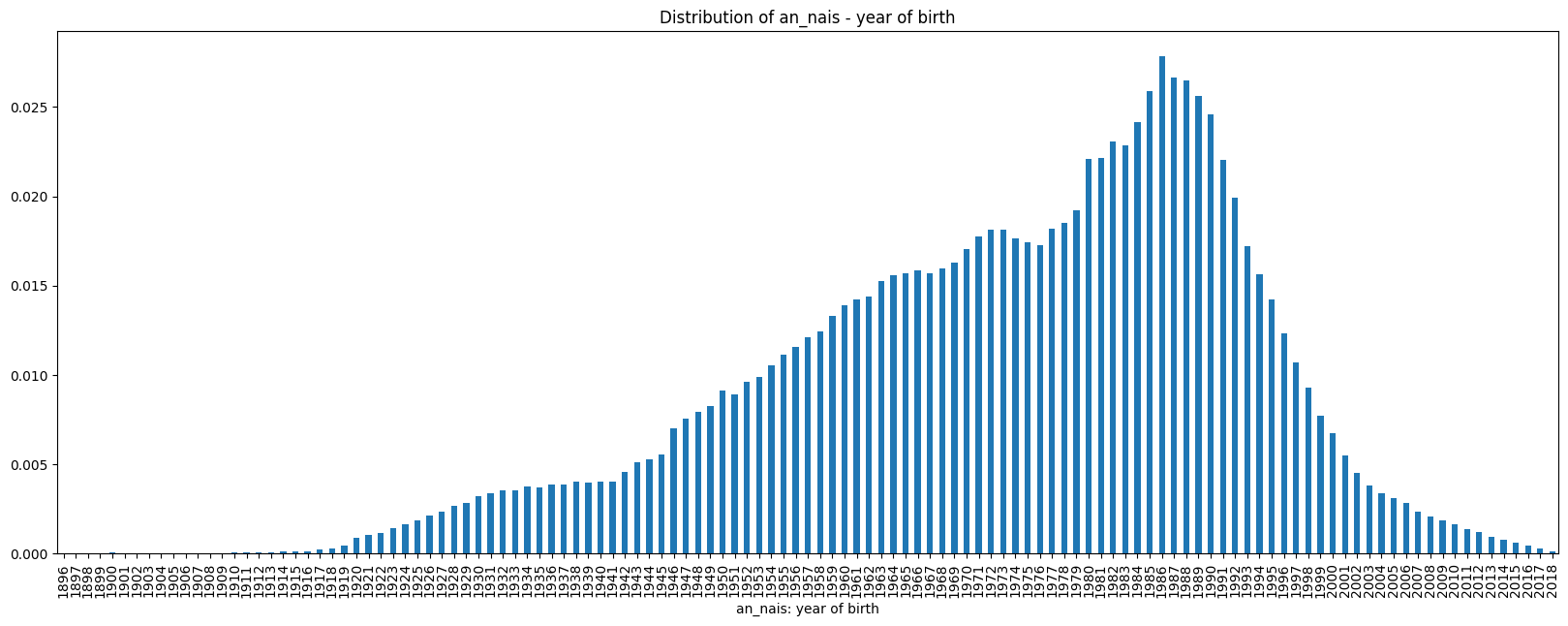


#### Column an\_nais (year of birth)



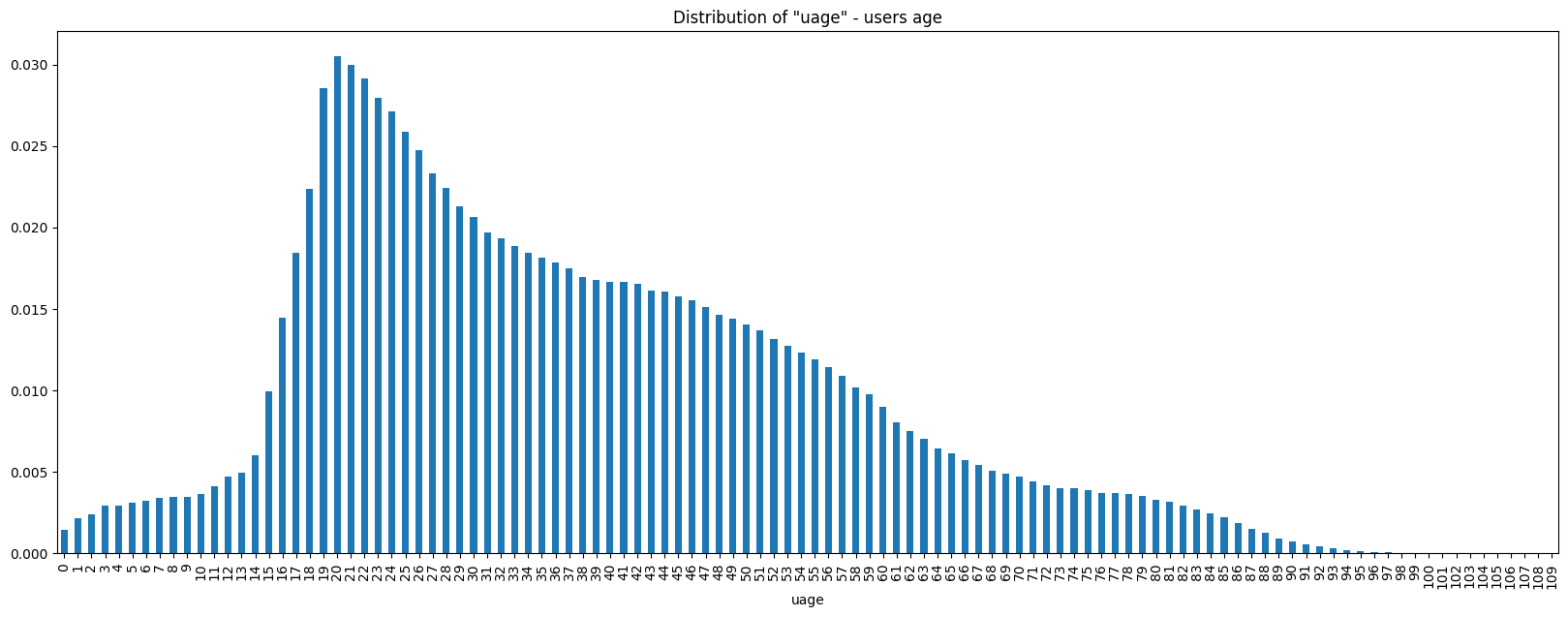
There are 2,418 rows with no year of birth. We replace them with the mode:





Create New Column User Age: uage = year of accident - year of birth.





Drop column an\_nais (year of birth in favor of new column user age: uage)



Find the optimal split for uage with respect to multiclass target grav with the help of “optbinning” package: (htps://gnpalencia.org/optbinning/tutorials/tutorial\_multiclass.html)

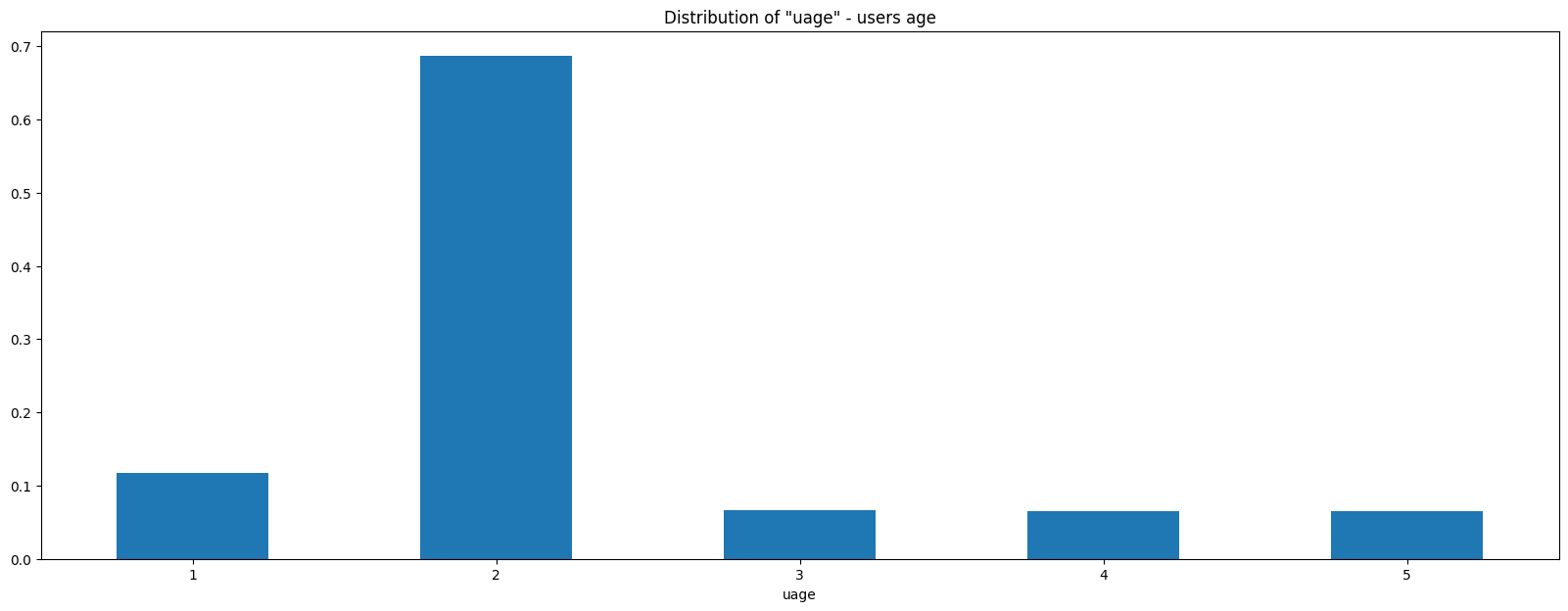


OPTIMAL

[18.5 53.5 59.5 69.5]

Result for optimal binning: Recode uage into 5 categories:

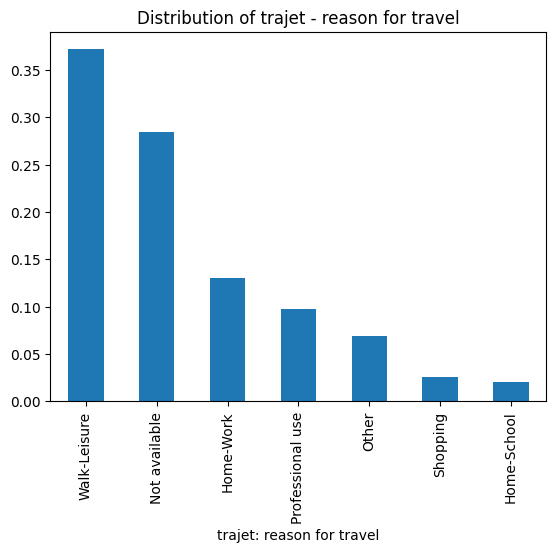




Final\_DF: uage will be replaced by 5 dummies, one for every uage\_category, counting the users per category.

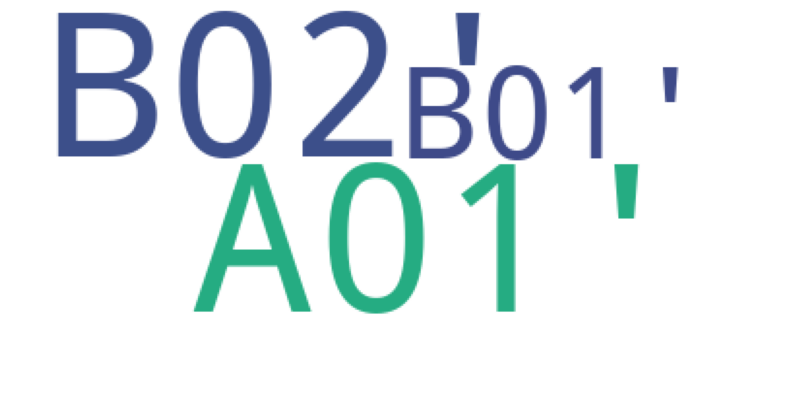
#### Column trajet (route, reason for travel)

Replace NaN with 0 (Not available).



Final\_DF: tarjet will be replaced by 7 dummies, one for every tarjet\_category, counting the users per category.

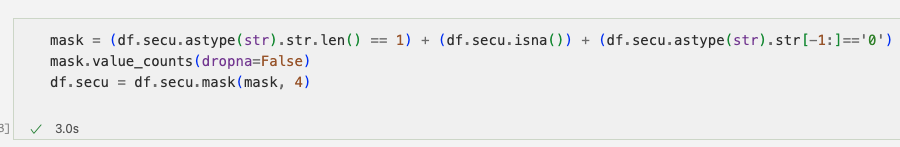
#### Column num\_veh (Vehicle ID)



Nothing to clean here.

Final\_DF: num\_veh will count the number of vehicles involved in the accident, i.e., the distinct num\_veh per Num\_Acc.

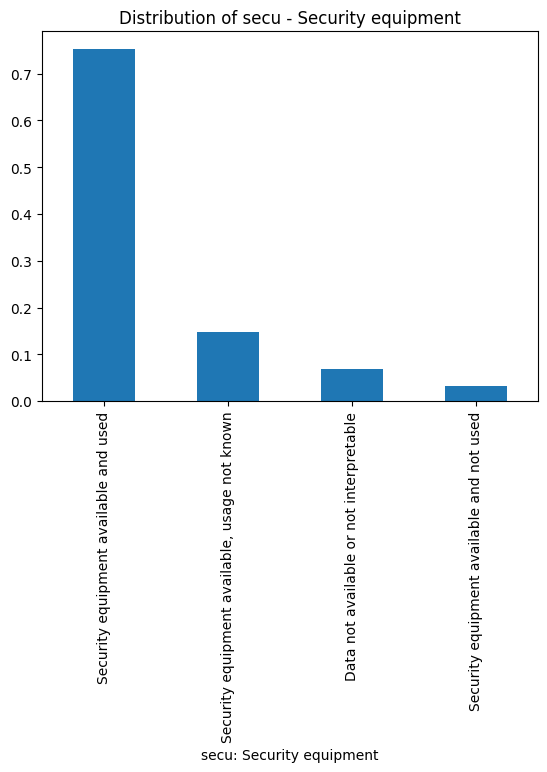
#### Colum secu - Safety equipment

Single digit numbers and numbers ending in 0 are not explained in the data-dictionary -> set to 4

Recode secu based on second digit:

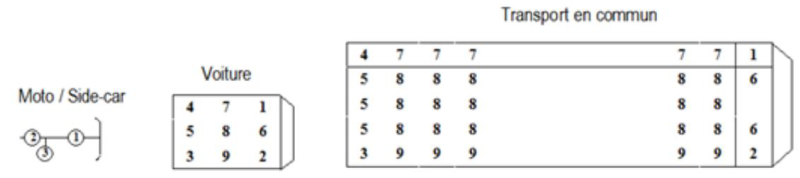
x1 -> 1 (Security equipment available and used)  
x2 -> 2 (Security equipment available and not used)  
x3 -> 3 (Security equipment available, usage not known)  
NAN -> 4 (data not available or not interpretable)





Final\_DF: secu will be replaced by 4 dummies, one for every secu\_category, counting the users per category.

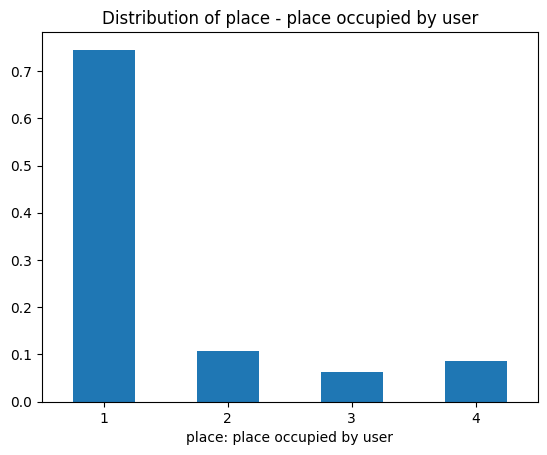
#### Column place (seat in car)



Cleaning: Replace undocumented 0 with NAN

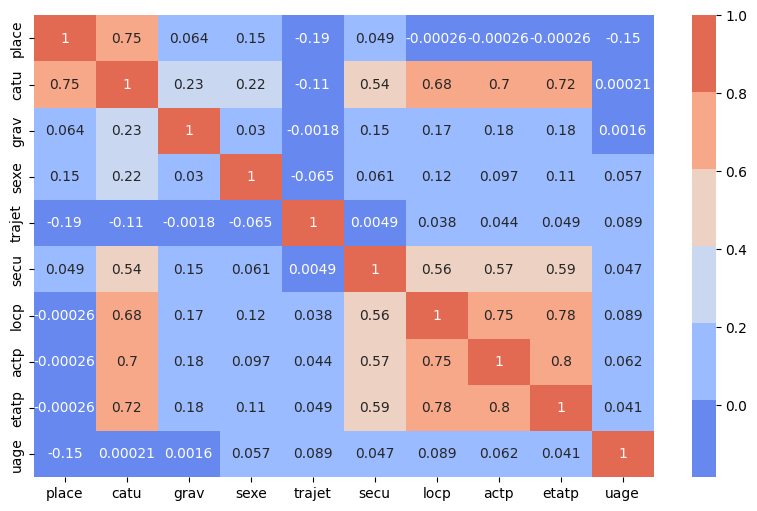
Create 4 Categories:

1 - driver (1)  
2 - passenger front seat (2,6)  
3 - passenger rear seat (3,4,5,7,8,9)  
4 - no info available



Final\_DF: place will be replaced by 4 dummies, one for every place\_category, counting the users per category.

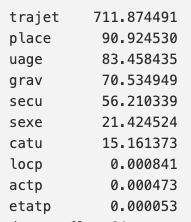
#### Correlations



Variable catu (user category, driver, passenger...) is strongly correlated to many other variables.

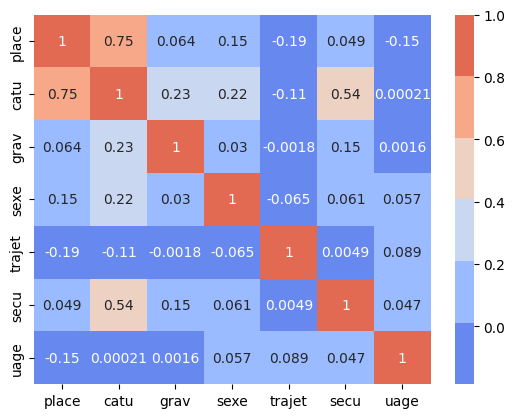
Checking the variance of the variables:



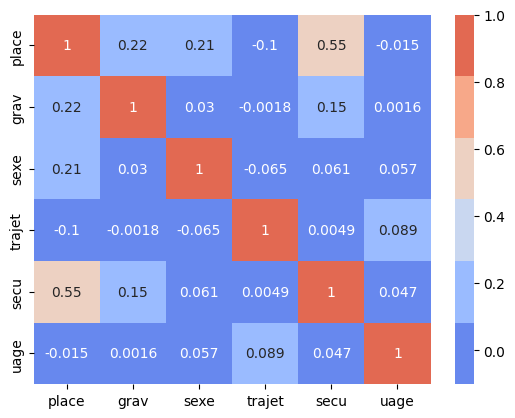


Variables with low variance are not expected to add much to the prediction of a model. Removing variables with lowest variances:





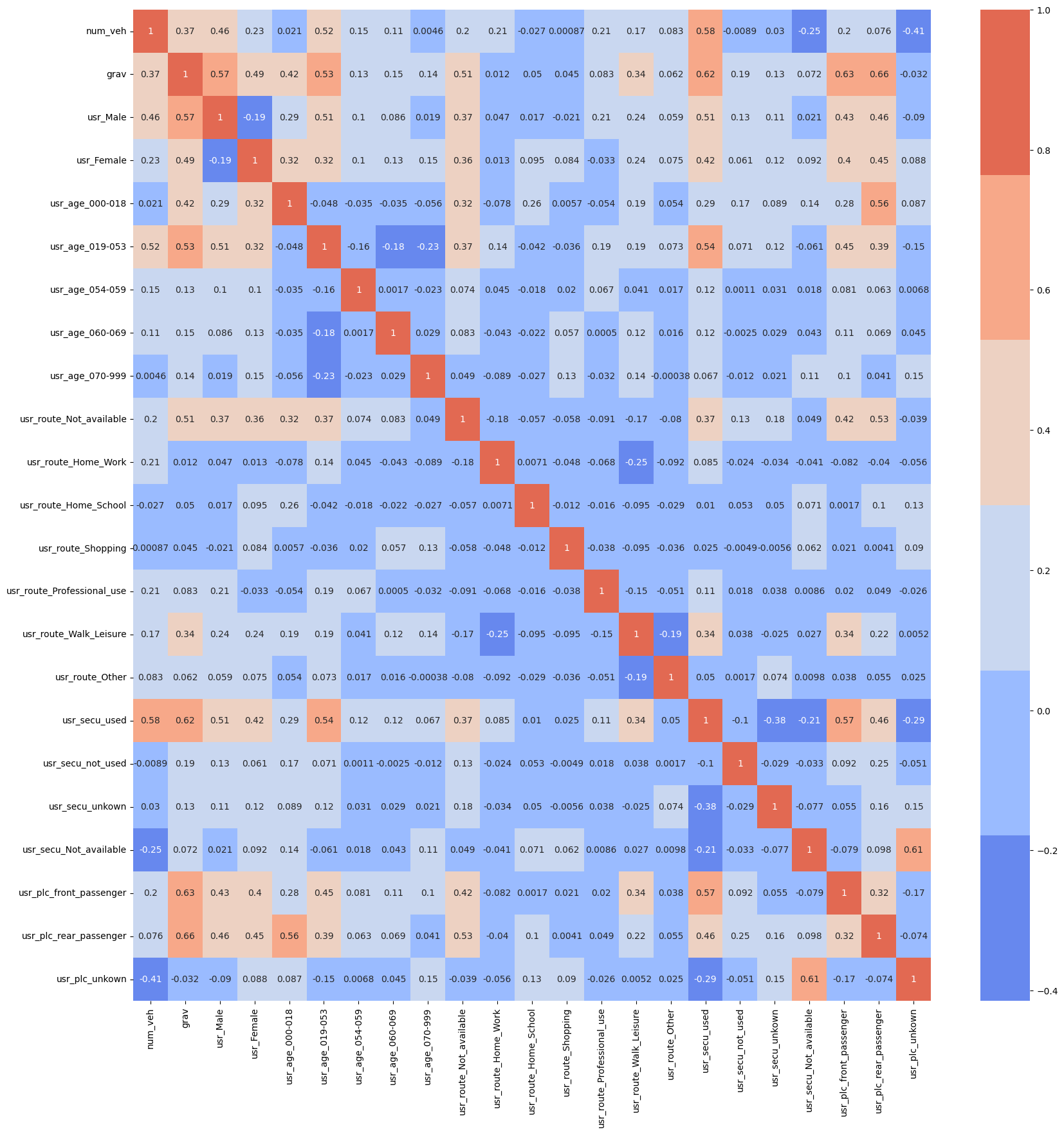
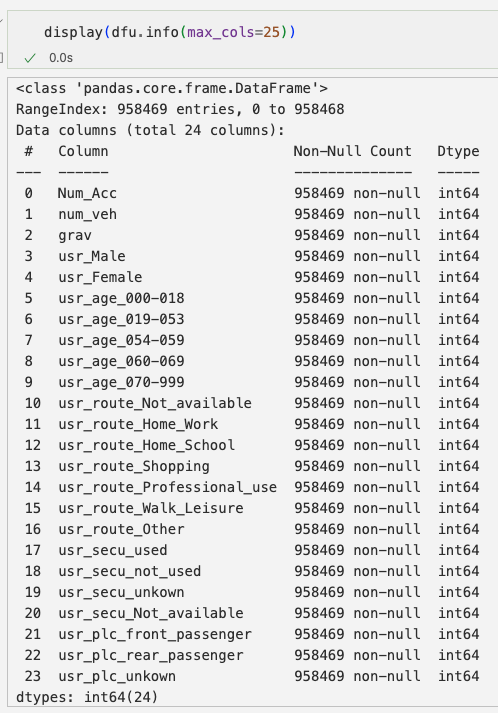
Drop catu, as this variable is strongly correlated to place, which we want to keep.



#### Widened Dataframe (1 row per Num\_Acc)

|  |
| --- |
| **col\_num\_veh** = df[['Num\_Acc', 'num\_veh']].groupby('Num\_Acc').num\_veh.nunique() |
| **col\_grav** = df[['Num\_Acc', 'grav']].groupby('Num\_Acc').sum() |
| col='sexe'  col\_short='sex'  actual\_col\_map = {  'sex\_1':'usr\_Male'  ,'sex\_2':'usr\_Female'}  col\_sex = df[['Num\_Acc', col]].join(pd.get\_dummies(df[col], prefix=col\_short, dtype="int64")).drop(col, axis=1).groupby('Num\_Acc').sum()  **col\_sex**=col\_sex.rename(columns=actual\_col\_map) |
| col='uage'  col\_short='uage'  actual\_col\_map = {  'uage\_1':'usr\_age\_000-018'  ,'uage\_2':'usr\_age\_019-053'  ,'uage\_3':'usr\_age\_054-059'  ,'uage\_4':'usr\_age\_060-069'  ,'uage\_5':'usr\_age\_070-999'  }  col\_uage = df[['Num\_Acc', col]].join(pd.get\_dummies(df[col], prefix=col\_short, dtype="int64")).drop(col, axis=1).groupby('Num\_Acc').sum()  **col\_uage** = col\_uage.rename(columns=actual\_col\_map) |
| col='trajet'  col\_short='traj'  actual\_col\_map = {  'traj\_0':"usr\_route\_Not\_available"  ,'traj\_1':"usr\_route\_Home\_Work"  ,'traj\_2':"usr\_route\_Home\_School"  ,'traj\_3':"usr\_route\_Shopping"  ,'traj\_4':"usr\_route\_Professional\_use"  ,'traj\_5':"usr\_route\_Walk\_Leisure"  ,'traj\_9':"usr\_route\_Other"  ,  }  col\_trajet = df[['Num\_Acc', col]].join(pd.get\_dummies(df[col], prefix=col\_short, dtype="int64")).drop(col, axis=1).groupby('Num\_Acc').sum()  **col\_trajet** = col\_trajet.rename(columns=actual\_col\_map) |
| col='secu'  col\_short='secu'  actual\_col\_map = {  'secu\_1': "usr\_secu\_used"  ,'secu\_2': "usr\_secu\_not\_used"  ,'secu\_3': "usr\_secu\_unkown"  ,'secu\_4': "usr\_secu\_Not\_available"  ,  }  col\_secu = df[['Num\_Acc', col]].join(pd.get\_dummies(df[col], prefix=col\_short, dtype="int64")).drop(col, axis=1).groupby('Num\_Acc').sum()  **col\_secu** = col\_secu.rename(columns=actual\_col\_map) |
| col='place'  col\_short='place'  actual\_col\_map = {  'place\_1': "usr\_plc\_driver"  ,'place\_2': "usr\_plc\_front\_passenger"  ,'place\_3': "usr\_plc\_rear\_passenger"  ,'place\_4': "usr\_plc\_unkown"  ,  }  col\_place = df[['Num\_Acc', col]].join(pd.get\_dummies(df[col], prefix=col\_short, dtype="int64", drop\_first=True)).drop(col, axis=1).groupby('Num\_Acc').sum()  **col\_place** = col\_place.rename(columns=actual\_col\_map) |
|  |
| **dfu** = pd.DataFrame(df.Num\_Acc.unique(), columns=['Num\_Acc']) |
| dfu=dfu.join(col\_num\_veh, on='Num\_Acc')  dfu=dfu.join(col\_grav, on='Num\_Acc')  dfu=dfu.join(col\_sex, on='Num\_Acc')  dfu=dfu.join(col\_uage, on='Num\_Acc')  dfu=dfu.join(col\_trajet, on='Num\_Acc')  dfu=dfu.join(col\_secu, on='Num\_Acc')  dfu=dfu.join(col\_place, on='Num\_Acc') |

Taking a look at the correlation map of the final dataframe for the users:

Saving final dataframe as pickle-file with xz-compression:



### Characteristics file

Caracteristics file describes the general circumstances of the accident.

The percentage of missing data is described in following prospect:

Missing data per column in %:

Num\_Acc 0.00  
an 0.00  
mois 0.00  
jour 0.00  
hrmn 0.00  
lum 0.00  
agg 0.00  
int 0.00  
atm 0.01  
col 0.00  
com 0.00  
adr 12.28  
gps 59.21  
lat 41.39  
long 41.39  
dep 0.00

Highest percentage of missing values is related to geo localization. These columns have been kept anyway in order not to prevent any possible geographical distribution analysis.

For columns “atm” (atmospheric conditions) and “col” (collision type) missing values have been replaced with “-1”

The presence of “zero” values has also been checked. For column “int” zero has been replaced with “-1”.

For other columns, like hour, zero value is meaningful and therefore it has not been replaced.

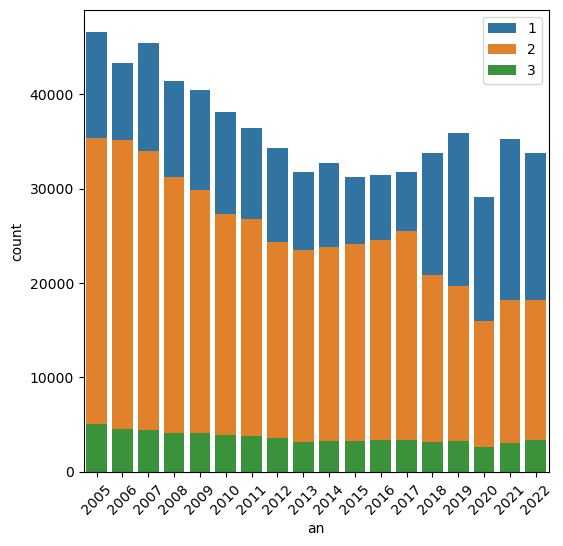
No duplicates are present in caracteristics dataframe.

* Column “an” - Year:

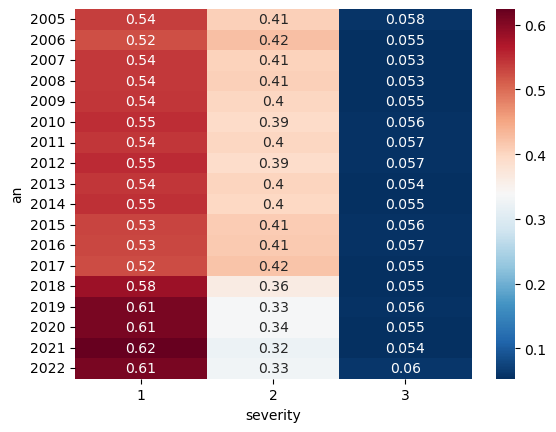
Column “an” has been harmonized in format “YYYY”.

The evolution of number of accidents has decreased significantly between 2005 and 2013.

From 2013 onwards it has slightly increased, especially for accidents with lower severity. The effects of the lockdown in 2020 can be spot in the chart.



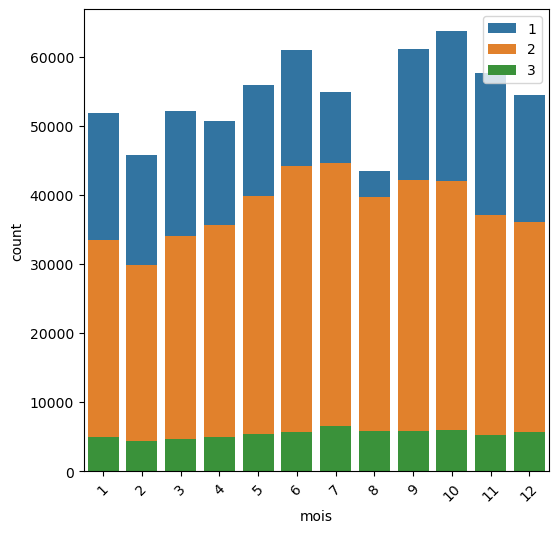
Looking at the split per year and severity, higher severity ratio looks stable along years, while claims having lower severity are increasing in the latest years.



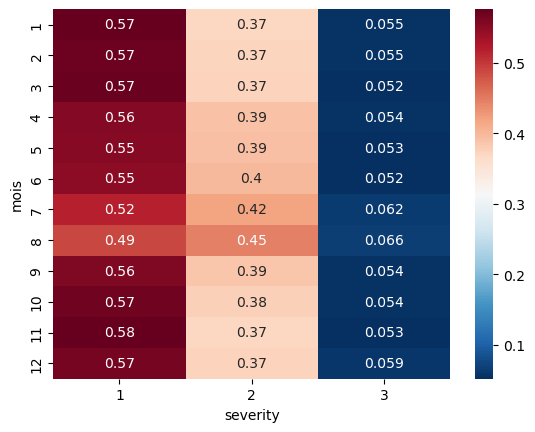
* Column “mois” - Month:

Distribution of accidents by month shows a seasonality. In particular, in the summer months there’s a reducion of number of accidents, together with a higher number of claims with high severity.

The drop of low severity claims in August can be justified by the lower traffic in the cities.

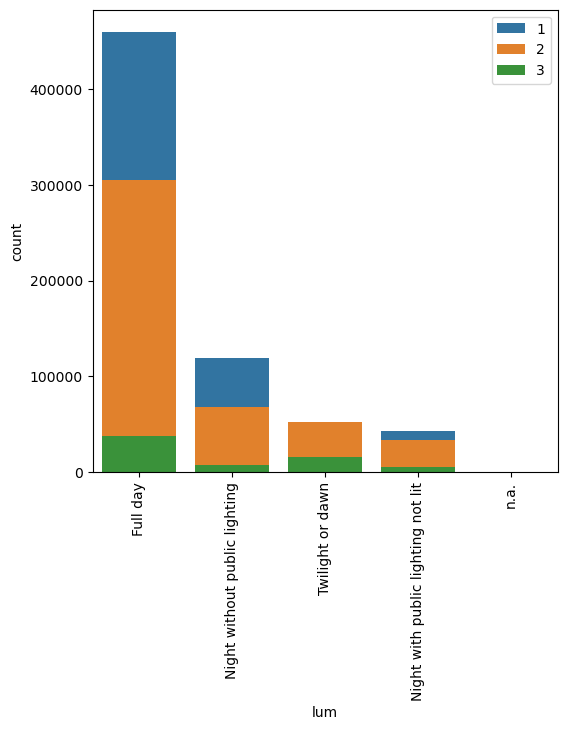


The incidence of high severity accidents can be spot in the following chart. It can be observed a higher incidence of high severity accidents in July and August.

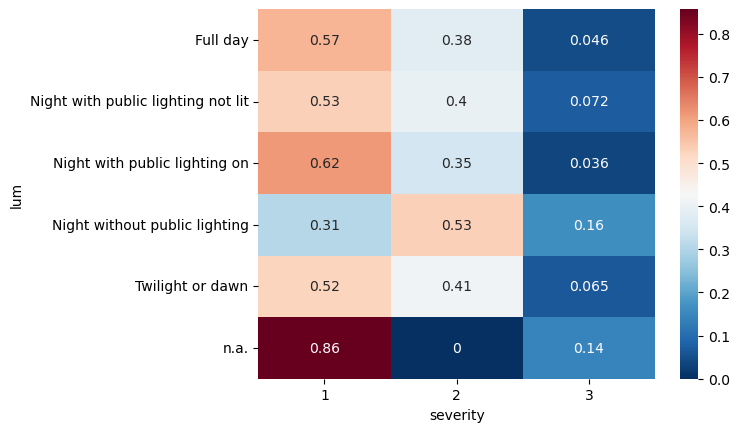


* Column “lum” - Light conditions:

Column “lum” refers to light conditions. The majority of accidents occurs during the day.



It is important to remark that accidents occurred during the night WITHOUT public lighting have an incidence of high severity 4 times higher than those occurred during the night WITH public lighting.



According to the results above the variable has been renamed as follows:

1: 'Full day': stays the same

2: 'Twilight or dawn': incorporate in '4: others'

3: 'Night without public lighting': stays the same (2: 'Night without public lighting')

4: 'Night with public lighting not lit': incorporate in '4: others'

5: 'Night with public lighting on': stays the same (3: 'Night with public lighting on')

New values:

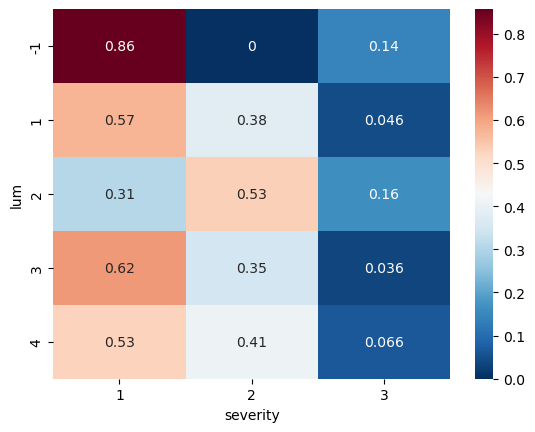
1: 'Full day'

2: 'Night without public lighting'

3: 'Night with public lighting on'

4: 'Others': includes 'Twilight or dawn' and 'Night with public lighting not lit'

We check again the distribution of the different values with our target variable:

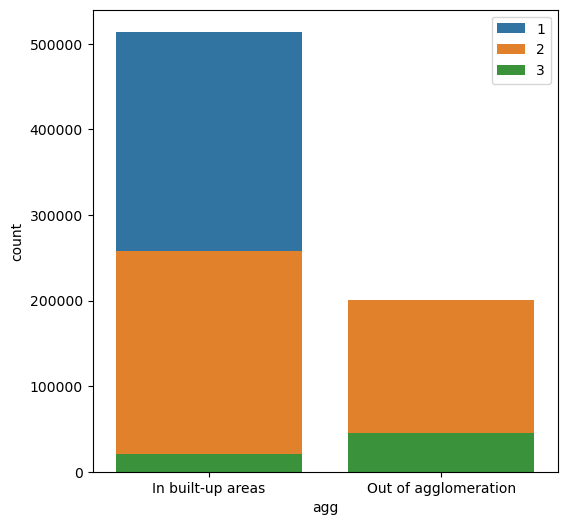


* Column “agg” - Location:

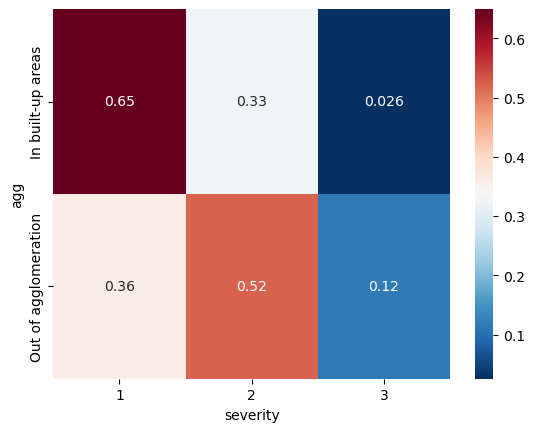
Column “agg” refers to the location of the accident. We can observe 2 possibile values:

* In built-up areas
* Out of agglomeration

Accidents occurred in built-up areas are more than twice the other class but with a lower severity distribution. Accidents occurred out of agglomeration have a higher severity and a higher mortality.



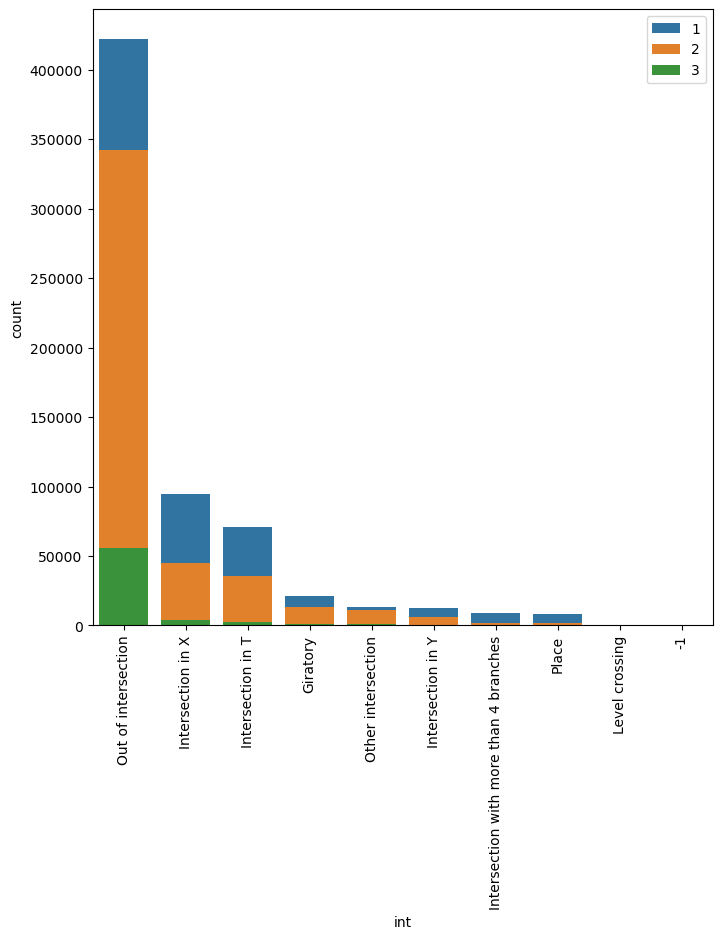
For out of agglomeration class the incidence of high severity levels is almost 5 times higher than in-built-up areas.



* Column “int” - Intersection:

Column “int” in which kind of intersection the accident occurred.

The majority of accidents occurred out of any intersection. In this class we have almost all accidents with high severity.





According to the results above the variable has been renamed as follows:

1: 'Out of intersection': stays the same

2: 'Intersection in X': grouped in '2: Intersection/other'

3: 'Intersection in T': grouped in '2: Intersection/other'

4: 'Intersection in Y': grouped in '2: Intersection/other'

5: 'Intersection with more than 4 branches': grouped in '2: Intersection/other'

6: 'Giratory': grouped in '2: Intersection/other'

7: 'Place': grouped in '2: Intersection/other'

8: 'Level crossing': stays the same

9: 'Other intersection': grouped in '2: Intersection/other'

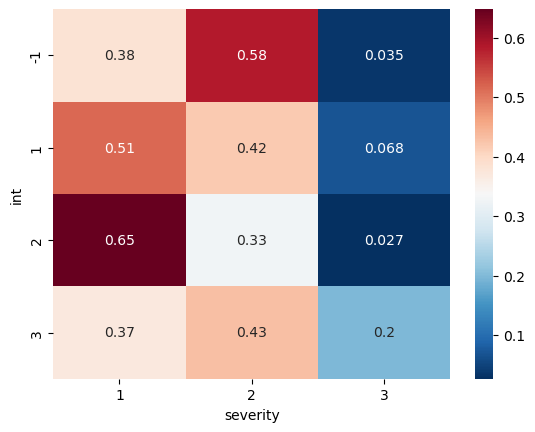
New values:

1: 'Out of intersection'

2: 'Intersection/other'

3: 'Level crossing'

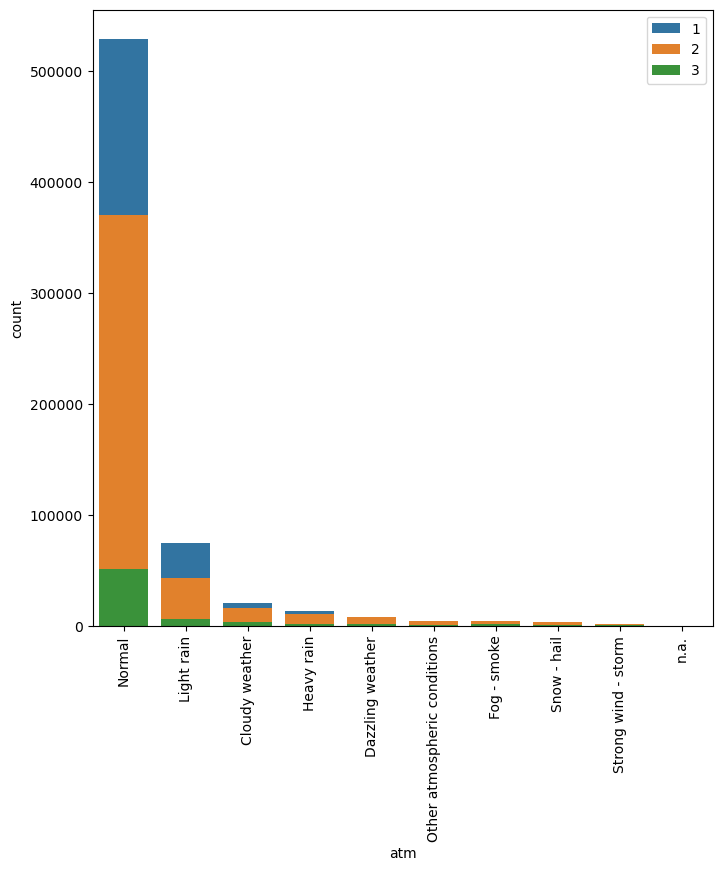
We check again the distribution of the different values with our target variable:



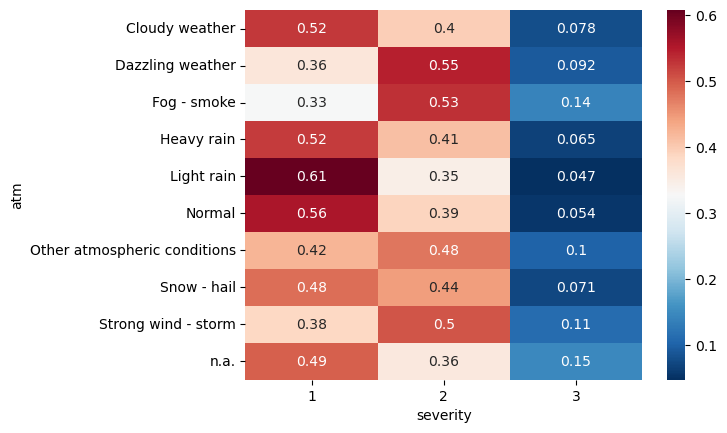
* Column “atm” - Atmospheric conditions:

Column “atm” specifies in which atmospheric conditions the accident occurred.

Most part of accidents occurred in normal conditions.



However, high severity incidence is 3 times higher in case of fog-smoke and almost twice in case of dazzling or cloudy weather.



According to the results above the variable has been renamed as follows:

1 - Normal: grouped into '1: good conditions: Normal/cloudy/light rain'

2 - Light rain: grouped into '1: good conditions: Normal/cloudy/light rain'

3 - Heavy rain: grouped into '2: medium conditions: heavy rain/snow/hail'

4 - Snow - hail: grouped into '2: medium conditions: heavy rain/snow/hail'

5 - Fog - smoke: grouped into '3: bad conditions/visibility: fog/smoke/storm/dazzling'

6 - Strong wind - storm: grouped into '3: bad conditions/visibility: fog/smoke/storm/dazzling'

7 - Dazzling weather: grouped into '3: bad conditions/visibility: fog/smoke/storm/dazzling'

8 - Cloudy weather: grouped into '1: good conditions: Normal/cloudy/light rain'

9 - Other: goes into '4: other'

New values:

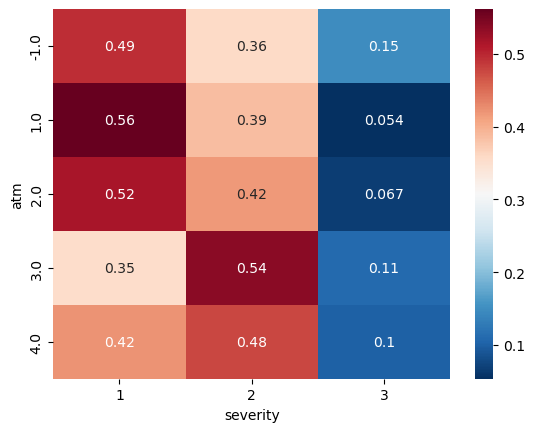
1: good conditions: Normal/cloudy/light rain

2: medium conditions: heavy rain/snow/hail

3: bad conditions/visibility: fog/smoke/storm/dazzling

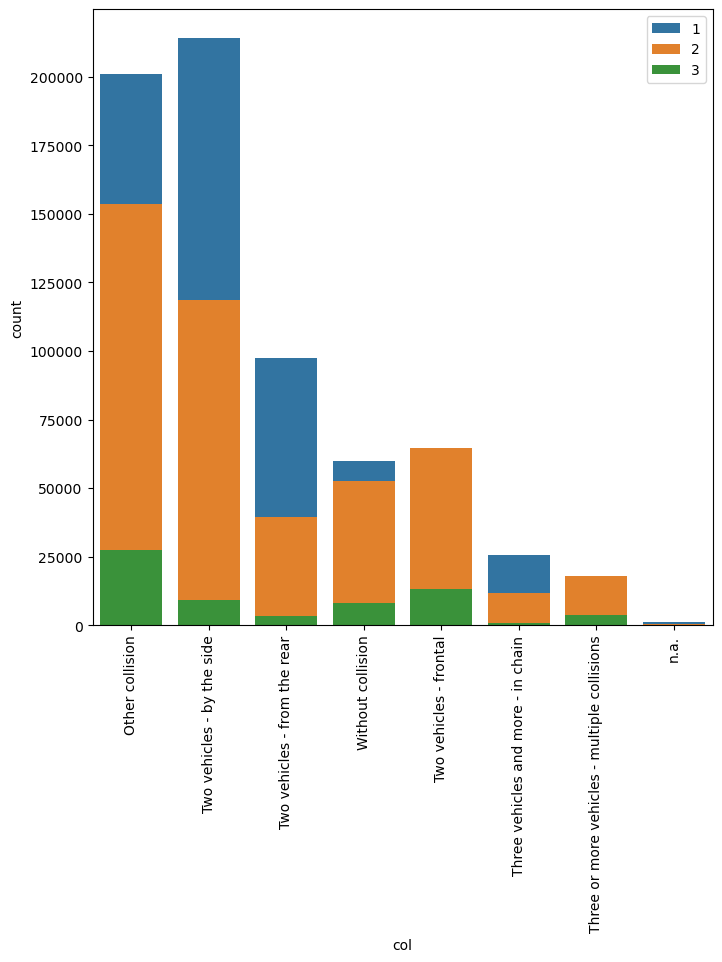
4: other

We check again the distribution of the different values with our target variable:

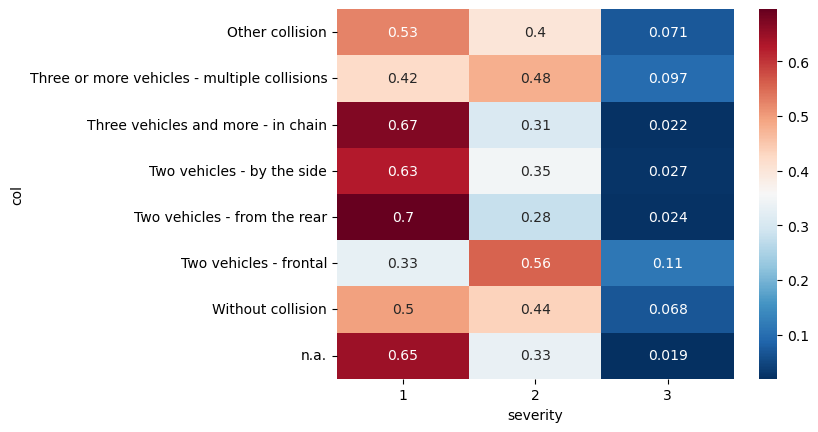


* Column “col” - Collision type:

Collision type shows how most part of accidents occurs side by side.



However, high severity accidents is 5 times higher in case of frontal or multiple collisions.



According to the results above the variable has been renamed as follows:

1 - Two vehicles - frontal: grouped into '1: frontal/multiple collisions'

2 - Two vehicles - from the rear: grouped into '2: rear/side/chain'

3 - Two vehicles - by the side: grouped into '2: rear/side/chain'

4 - Three vehicles and more - in chain: grouped into '2: rear/side/chain'

5 - Three or more vehicles - multiple collisions: grouped into '1: frontal/multiple collisions'

6 - Other collision: grouped into '3: without collision/other'

7 - Without collision: grouped into '3: without collision/other'

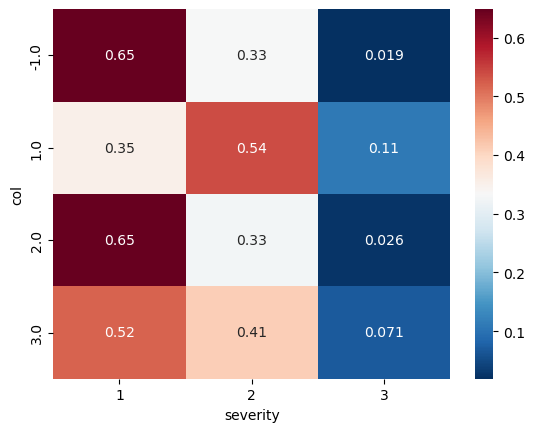
New values:

1: frontal/multiple collisions

2: rear/side/chain

3: without collision/other

We check again the distribution of the different values with our target variable:



* Column “dep” - Department:

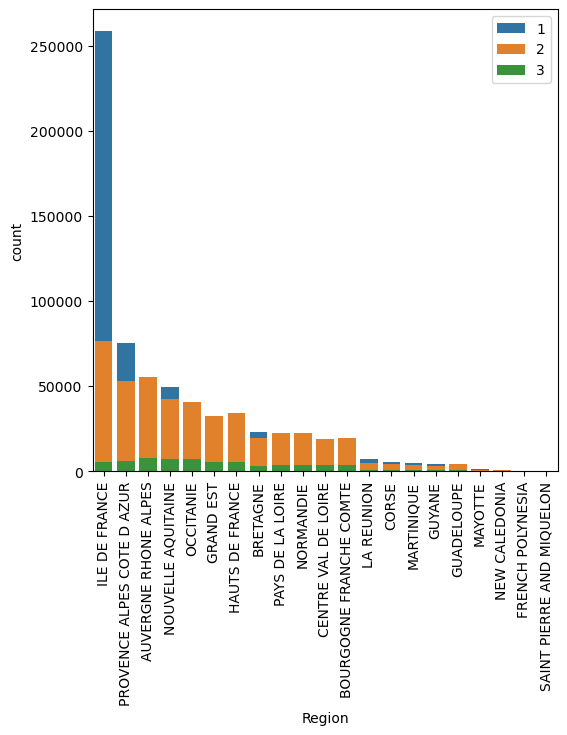
Department refers to the location of the accident.

As the number of departments in France is about 100, they have been grouped by Region according the following Table:

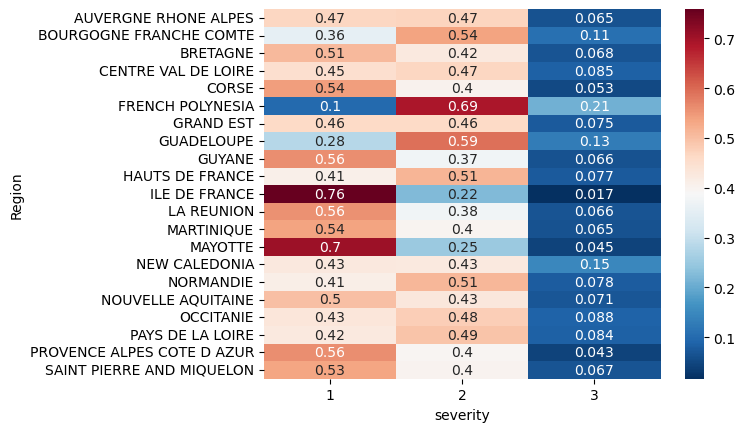
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Region\_ID | Region | Region\_ID | Region |  |
| 1 | GUADELOUPE | 52 | PAYS DE LA LOIRE |  |
| 2 | MARTINIQUE | 53 | BRETAGNE |  |
| 3 | GUYANE | 75 | NOUVELLE AQUITAINE |  |
| 4 | LA REUNION | 76 | OCCITANIE |  |
| 6 | MAYOTTE | 84 | AUVERGNE RHONE ALPES |  |
| 11 | ILE DE FRANCE | 93 | PROVENCE ALPES COTE D AZUR |  |
| 24 | CENTRE VAL DE LOIRE | 94 | CORSE |  |
| 27 | BOURGOGNE FRANCHE COMTE | 991 | SAINT PIERRE AND MIQUELON |  |
| 28 | NORMANDIE | 992 | FRENCH POLYNESIA |  |
| 32 | HAUTS DE FRANCE | 993 | NEW CALEDONIA |  |
| 44 | GRAND EST | -1 | NA |  |
|  |  |  |  |  |

This manipulation allows to work on a lower number of values.

According to this new variable we can classify the distribution of severity of accidents.

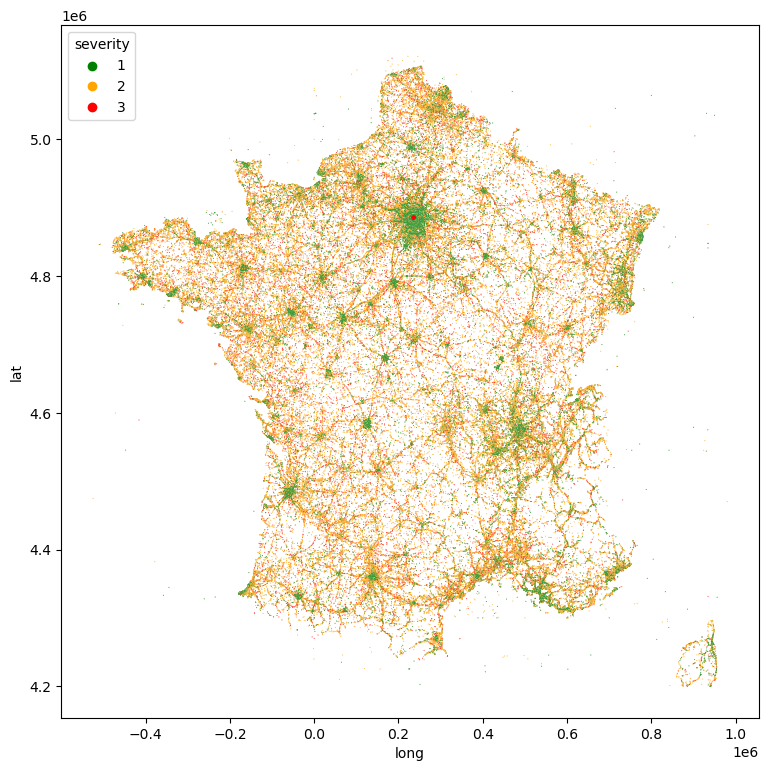


We can observe most “low severity” claims occur in “Ile de France” Region, i.e. Paris region.



* Columns “lat”-”long” - Latitude Longitude:

Latitude and longitude coordinates enable us to observe how accidents occurred in in-built areas are more often with lower severity. High severity claims occur outside the cities.



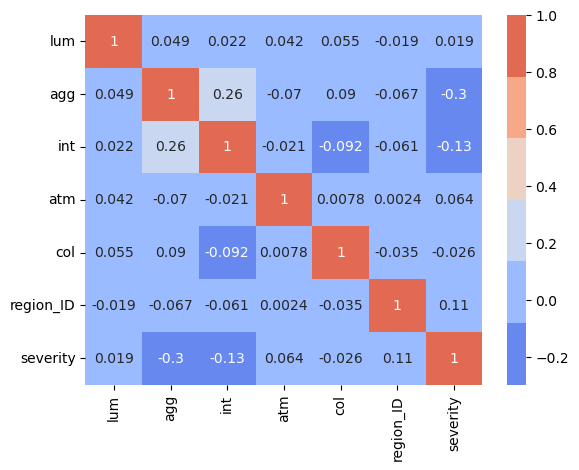
The map confirms the results coming from the analysis per Region. Most low-severity accidents are concentrated in Paris area.

Correlation coefficient with target:

We then checked the correlation coefficient of our categorical variables with our target variable “severity”.

The considered variables are:

* Lum - Light conditions
* Agg - Location
* Int - Intersection
* Atm - Atmospheric conditions
* Col - Collision type
* Region\_ID – Region



A medium correlation intensity can be observed for variables “agg-Location”, “int-Intersection” and “region\_ID”.

* Connection analysis

Qualitative variable have been analysed in order to understand the level of connection with the severity variable.

The used test is Cramer together with the related p-value.

The test results are reported here below:

target and lum - cramers\_v: 0.133 - p-value: 0.0

target and agg - cramers\_v: 0.297 - p-value: 0.0

target and int - cramers\_v: 0.105 - p-value: 0.0

target and atm - cramers\_v: 0.057 - p-value: 0.0

target and col - cramers\_v: 0.156 - p-value: 0.0

target and com - cramers\_v: 0.297 - p-value: 0.0

target and dep - cramers\_v: 0.272 - p-value: 0.0

target and Region - cramers\_v: 0.203 - p-value: 0.0

A medium connection can be observed for column “agg” – Location – and “col” – Collision type, as well as geographical variables like com, dep and region.

* Random Forest Classifier – feature importance

The outcomes of the previous analyses are confirmed also using the Random Forest Classifier algorithm.

