Machine Learning Project

on

Predicting Severity of Road Traffic Accident

in association with

alfatraining Bildungszentrum GmbH

Professor: Elisabeth Staegemann

Published on: 24.09.2021

Table of Contents

1. Particpants & Areas of Work

Priya Subramanian

(Main Focus: Visualisation, Report)

Felix Schulze

(Main Focus: ML-Algorithms)

Michael Pollich

(Main Focus: Data Preparation, Report)

2.Introduction

2.1 Aim

Traffic accidents are a significant source of deaths, injuries, property damage, and a major concern for public health and traffic safety. Accidents are also a major cause of traffic congestion and delay. Effective management of accident is crucial to mitigating accident impacts and improving traffic safety and transportation system efficiency. One of the major steps of the accident response team or program is „predicting severity“, of high importance.

The methods followed in this project varied from the traditional data cleaning, explanatory data analysis, statsitcal methods to sophisticated machine learning (ML) methods. The predictive ML models of the project contributes to generating crucial information, which can be used to adopt appropriate measures for reducing the aftermath of crashes. Moreover, accurate severity prediction of traffic crashes can help hospitals to provide medical care quickly if a crash occurs. It could also be used in a specific context by emergency services team or program to manage knowing more precisely the need of resources for the particular accident.

2.2 Hypothesis

Aus dem umfangreichen Datensatz lassen sich vor Beginn der Untersuchung viele Hypothesen ableiten, bei denen oft mehrere Variablen beteiligt sind. Z.B.

Die Schwere der Unfallfolgen sollte entscheidend durch die Fahrgeschwindigkeit beeinflusst werden. Je schneller gefahren wurde, desto wahrscheinlicher sind Todesfälle oder Krankenhauseinlieferungen. D.h. je höher die Geschwindigkeitsbegrenzung, desto wahrscheinlicher sind schwere Folgen. Auf Autobahnen und Bundesstraßen wird meist schneller gefahren als auf Stadtstraßen. Allerding sind Fußgänger und Radfahrer als ungeschützte Verkehrsteilnehmer vorwiegend innerhalb geschlossener Ortschaften anzutreffen. In großen Städten ist außerdem die Verkehrsdichte hoch und der Verkehr unübersichtlich, was zu mehr Unfällen führt.

Je näher ein Notruftelefon zur Unfallstelle ist, desto schneller kann Hilfe geholt werden und desto weniger Todesfälle gibt es.

Je ungünstiger die Wetterbedingungen (Licht, Straßenzustand ,Wetter), desto häufiger sind schwere Unfallfolgen.

Da Autos und Sicherheitsequipment vorwiegend für Männer optimiert sind, kann vermutet werden, dass Frauen häufiger schwere Unfallfolgen erleiden als Männer.

Je älter ein Unfallbeteiligter ist, desto wahrscheinlicher sind schwere Unfallfolgen

…

2.3 Quality criteria

The project aims to study the prediction accuracies of models and aim to achieve atleast 50% accuracies.

Accuracy good, when about equal values for false positives and false negatives

With more time we could also include the following:

Precision: penalises for false positives

Recall: penalises for false negatives

F1-score: penalises for both false negatives and false positives

2.4 Machine Learning techniques

Classification Model using Supervised learning with categorical data

Wir haben uns für folgende Algorithmen entschieden:

KNN: langsam, berücksichtigt die Interaktion zwischen den Daten gut

Gausian Naive Bayes: Features sind relativ unabhängig, sehr schnell

Random Forest: nicht so sensibel für unvorbereitete Daten, benutzt viele unterschiedliche Bäume

3.Data Description

3.1 Data Source

The data for this project is collected from the transport department of French Government. The Open source Data provides overall road accidents occured for the year 2019.

link: https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2019/

3.2 Data Descritpion & Contents

Size : 21,08 MB

There are 4 datasets in this challenge, which represents as below:

Column Description for all 4 datasets

Dataset Name	No. Of rows	No.of Columns	Description
caracteristiques_2019.csv	58840	15	basic data about the accidents and it's characteristics
lieux_2019.csv	58840	18	data about the location of the accidents
usagers_2019.csv	100710	11	data about the people involved in the accidents
vehicules_2019.csv	132977	15	data about the vehicles involved

Original Col_name	Actual Col_name	Description
Num_Acc	ID_Acc	Identification number of the accident.
jour	Day	Day of the accident.
mois	Month	Month of the accident.
an	Year	Year of accident.
hrmn	Time	Hour and minutes of the accident. This one is tricky, it correspond to a percentage of 24h (your turn to convert it into day time)
lum	Light	Light: lighting conditions in which the accident occurred
		1. Full day
		2. Twilight or dawn
		3. Night without public lighting
		4. Night with public lighting not on
		5. Night with public lighting on
com	Municaplity	Municipality: The municipality number is a code given by INSEE. The code is made up of the code INSEE of the department followed by 3 digits.
agg	Location	Location :
		1. Outside agglomeration
		2. In built-up areas
int	int	Intersection:
		1. Excluding intersection
		2. Intersection in X
		3. T-intersection
		4. Y intersection
		5. Intersection with more than 4 branches
		6. Roundabout
		7. Place
		8. Level crossing
		9. Other intersection
atm	atm	Atmospheric conditions:
		1. Normal
		2. Light rain
		3. Heavy rain
		4. Snow. hail
		5. Fog. smoke
		6. Strong wind, storm
		7. Dazzling weather
		8. Cloudy weather
		9. Other
col	col	Collision type:
		1. Two vehicles. frontal
		2. Two vehicles. from the rear
		3. Two vehicles. from the side
		4. Three vehicles and more. in a chain
		5. Three or more vehicles. multiple collisions
		6. Other collision 7. No collision
adr	adr	Postal address: variable entered for accidents occurring in built-up areas.
lat	lat	latitude
long	long	longitude

Original Col_name	Actual Col_name	Description
catr	road_category	Road category:
		1. Highway
		2. National road
		3. Departmental road
		4. Communal roads
		5. Outside the public network
		6. Parking lot open to public traffic
		7. Urban metropolis roads
		8. other
voie	route_num	Route Number
V1	road_index1	Numerical index of the road number (example: 2 bis, 3 ter etc.).
V2	road_index2	Alphanumeric road index letter.
circ	traffic_regime	Traffic regime:
		1. One way
		2. Bidirectional
		3. A separate carriageway
		4. With variable assignment channels
nbv	traffic_lanes	Total number of traffic lanes.
vosp	reserved_lane	Indicates the existence of a reserved lane, regardless of whether or
		not the accident took place on this way.
		1. Not applicable
		2. Cycle path
		3. Cycle lane
		4. Reserved lane
prof	prof	Longitudinal profile describes the gradient of the road at the location of the accident:
		1. Flat
		2. Slope
		3. hilltop
		4. Bottom of coast
pr	upstream_terminal	Number of the associated PR (number of the upstream terminal). The value -1 means that the PR is not informed.
pr1	distance_to_upstream	Distance in meters from the PR (in relation to the upstream terminal). The value -1 means that the PR is not informed.
plan		Plan layout:
		1. rectilinear part
		2. In a curve to the left
		3. In a curve to the right 4. In "S"
lartpc	central_reservation_width	Width of the central reservation (TPC) if it exists (in m).
larrout	roadwidth_vehicles	Width of the roadway used for vehicular traffic is not included in the stopping strips emergency, TPC and parking spaces (in m).
surf	surf	Surface condition:

		1. Normal
		2. Wet
		3. Puddles
		4. Flooded
		5. Snowy
		6. Mud
		7. Icy
		8. Fat. oil
		9. Other
infra	infra	Development. Infrastructure:
		1. None
		2. Underground. tunnel
		3. Bridge. flyover
		4. Exchanger or connection sling
		5. Railroad
		6. Crossroads
		7. Pedestrian zone
		8. Toll zone
		9. Site
		10. Others
situ	situ	Situation of the accident:
		1. None
		2. On the road
		3. On emergency lane
		4. On the shoulder
		5. On the sidewalk
		6. On a cycle path
		7. On other special track
		8. Others
vma	speed_limit	Maximum authorized speed at the scene and at the time of the accident.

Original Col_name	Actual Col_name	Description
vehicle_id	ID_veh	Unique identifier of the vehicle used for each user occupying this vehicle (including pedestrians who are attached to the vehicles which collided with them). Numerical code.
Num_veh	Vehicle_No	Identifier of the vehicle taken back for each of the users occupying this vehicle (including pedestrians who are attached to the vehicles which collided with them). Alphanumeric code.
senc	Flow_Direction	Flow direction :
		1. Unknown
		2. PK or PR or increasing postal address number
		3. PK or PR or decreasing postal address number
		4. Lack of reference
catv	Vehicle_category	Vehicle category:
		1. Not determinable
		2. Bicycle
		3. Moped <50cm3
		5. Reference not used since 2006 (registered scooter)
		6. Reference unused since 2006 (motorcycle)
		7. Reference unused since 2006 (sidecar)
		8. VL only
		9. Reference unused since 2006 (VL + caravan)
		10. Reference not used since 2006 (light vehicles + trailer)
		11. VU only 1.5T <= PTAC <= 3.5T with or without trailer (formerly VU only 1.5T <= PTAC <= 3.5T)
		12. Reference not used since 2006 (VU (10) + caravan)
		13. Reference not used since 2006 (VU (10) + trailer)
		14. PL only 3.5T <PTCA <= 7,5T
		15. PL only > 7.5T
		16. PL> 3,5T + trailer
		17. Road tractor only
		18. Road tractor + semi-trailer
		19. Reference not used since 2006 (public transport)
		20. Reference not used since 2006 (tram)
		21. Special gear
		22. Farm tractor
		23. Scooter <50 cm3
		24. Motorcycle> 50 cm3 and <= 125 cm3
		25. Scooter> 50 cm3 and <= 125 cm3
		26. Motorcycle> 125 cm3
		27. Scooter> 125 cm3
		28. Light quad <= 50 cm3 (Quadricycle without bodywork engine)
		29. Heavy quad> 50 cm3 (Quadricycle without bodywork engine)
		30. Bus
		31. Coach
		32. Train
		33. Tram
		34. 3WD <= 50 cm3

		35. 3WD> 50 cm3 <= 125 cm3
		36. 3WD> 125 cm3
		37. EDP with motor
		38. EDP without motor
		39. VAE
		40. Other vehicle
obs	Fix_Obst_st	Fixed obstacle struck:
		1. Not applicable
		2. Parked vehicle
		3. Tree
		4. Metal slide
		5. Concrete slide
		6. Other slide
		7. Building, wall, bridge pier
		8. Vertical signage support or emergency call station
		9. Post
		10. Street furniture
		11. Parapet
		12. Island, refuge, upper terminal
		13. Sidewalk edge
		14. Ditch, embankment, rock face
		15. Other fixed obstacle on the road
		16. Other fixed obstacle on sidewalk or shoulder
		17. Clearance of the roadway without obstacle
		18. Nozzle. aqueduct head
obsm	Mov_Obst_st	Movable obstacle struck:
		1. None
		2. Pedestrian
		3. Vehicle
		4. Rail vehicle
		5. Domestic animal
		6. Wild animal
		7. Other
choc	Shock_pt	Initial shock point:
		1. None
		2. Before
		3. Right front
		4. Front left
		5. Rear
		6. Right back
		7. Left rear
		8. Right side
		9. Left side
		10. Multiple shocks (rolls)
manv	Maneuver	Main maneuver before the accident:
		1. Unknown

		2. Without change of direction
		3. Same direction, same row
		4. Between 2 lines
		5. In reverse
		6. In the wrong way
		7. Crossing the central reservation
		8. In the bus lane, in the same direction
		9. In the bus lane, in the opposite direction
		10. By inserting
		11. By making a U-turn on the road Changing lane
		12. Left
		13. Right Deported
		14. Left
		15. Right Turning
		16. Left
		17. Right Exceeding
		18. Left
		19. Right Various
		20. Crossing the road
		21. Parking maneuver
		22. Avoidance maneuver
		23. Door opening
		24. Stopped (except parking)
		25. Parked (with occupants
		26. Traveling on sidewalk
		27. Other maneuvers
motor	motor	Vehicle engine type:
		1. Unknown
		2. Hydrocarbons
		3. Electric hybrid
		4. Electric
		5. Hydrogen
		6. Human
		7. Other
occutc	No.of_occupants	Number of occupants in public transport.

Original Col_name	Actual Col_name	Description
id_vehicule	ID_veh	Unique identifier of the vehicle used for each user occupying this vehicle (including pedestrians who are attached to the vehicles which collided with them). Numerical code.
Num_veh	Vehicle_No	Identifier of the vehicle taken back for each of the users occupying this vehicle (including pedestrians who are attached to the vehicles which collided with them). Alphanumeric code.
place	place	Used to locate the space occupied in the vehicle by the user at the time of the accident Check on this link for the pattern : https://ibb.co/NsTxbXP
catu	User_category	User category:
		1. Driver
		2. Passenger
		3. Pedestrian
grav	Severity	victims plus unharmed:
		1. Unharmed
		2. Killed
		3. Injured hospitalized
sexe	Gender	Driver gender:
		1. Male
		2. Female
An_nais	DOB_driver	Year of birth of the driver
trajet	Reason_for_travel	Reason for travel at the time of the accident:
		1. Not specified
		2. Home. work
		3. Home. school
		4. Shopping. shopping
		5. Professional use
		6. Walk. leisure
		7. Other
secu1	Safety_equip_1	indicates the presence and use of safety equipment:
		1. No equipment
		2. Belt 12
		3. Helmet
		4. Children's device
		5. reflective vest
		6. Airbag (2WD / 3WD)
		7. Gloves (2WD / 3WD)
		8. Gloves + Airbag (2WD / 3WD)
		9. Not determinable
		10. Other
secu2	Safety_equip_2	indicates the presence and use of safety equipment:
		1. No equipment
		2. Belt
		3. Helmet
		4. Children's device
		5. reflective vest
		6. Airbag (2WD / 3WD)

		7. Gloves (2WD / 3WD)
		8. Gloves + Airbag (2WD / 3WD)
		9. Not determinable
		10. Other
secu3	Safety_equip_3	indicates the presence and use of safety equipment:
		1. No equipment
		2. Belt
		3. Helmet
		4. Children's device
		5. reflective vest
		6. Airbag (2WD / 3WD)
		7. Gloves (2WD / 3WD)
		8. Gloves + Airbag (2WD / 3WD)
		9. Not determinable
		10. Other
locp	Ped_loc	Pedestrian location:
		1. Not applicable On roadway:
		2. 50 m from the pedestrian crossing
		3.50 m from the pedestrian crossing On the pedestrian crossing:
		4. Without light signaling
		5. With various light signals:
		6. On the sidewalk
		7. On the shoulder
		8. On refuge or BAU
		9. On the side aisle
		10. Unknown
actp	Ped_act	Pedestrian action:
		1. Not specified or not applicable
		2. Direction of colliding vehicle
		3. Opposite direction of the vehicle Various
		4. Crossing
		5. Masked
		6. Playing. running
		7. With animal
		8. Other A. Get on / off the vehicle B. Unknown
etap	Ped_acc	This variable is used to specify whether the injured pedestrian was alone or not:
		1. Alone
		2. Accompanied
		3. In a group

Dataset Description for all 4 datasets

Vehicles:

```
<class 'pandas.core.frame.DataFrame'>
Index: 100710 entries, 138 306 524 to 137 982 130
Data columns (total 10 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   ID_Acc            100710 non-null  int64
 1   Vehicle_No        100710 non-null  object
 2   Flow_Direction    100710 non-null  int64
 3   Vehicle_category  100710 non-null  int64
 4   Fix_Obst_st       100710 non-null  int64
 5   Mov_Obst_st       100710 non-null  int64
 6   Shock_pt          100710 non-null  int64
 7   Maneuver          100710 non-null  int64
 8   motor             100710 non-null  int64
 9   No.of_occupants   892 non-null     float64
dtypes: float64(1), int64(8), object(1)
memory usage: 10.5+ MB
```

```
vehicles.describe()
```

	ID_Acc	Flow_Direction	Vehicle_category	Fix_Obst_st	Mov_Obst_st	Shock_pt	Maneuver	motor	No.of_occupants
count	1.007100e+05	100710.000000	100710.000000	100710.000000	100710.000000	100710.000000	100710.000000	100710.000000	892.000000
mean	2.019000e+11	1.557075	12.632817	0.990646	1.627098	2.926045	7.267401	1.171353	1.818386
std	1.697432e+04	0.833885	12.800755	3.088678	1.120068	2.421520	8.077319	1.112887	2.858931
min	2.019000e+11	-1.000000	0.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	0.000000
25%	2.019000e+11	1.000000	7.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
50%	2.019000e+11	1.000000	7.000000	0.000000	2.000000	2.000000	2.000000	1.000000	1.000000
75%	2.019000e+11	2.000000	10.000000	0.000000	2.000000	4.000000	15.000000	1.000000	1.000000
max	2.019001e+11	3.000000	99.000000	17.000000	9.000000	9.000000	26.000000	6.000000	33.000000

Correlations: no significant correlations

```
vehicles.corr()
```

	ID_Acc	Flow_Direction	Vehicle_category	Fix_Obst_st	Mov_Obst_st	Shock_pt	Maneuver	motor	No.of_occupants
ID_Acc	1.000000	-0.004331	0.004000	0.004541	0.001722	-0.000754	0.003393	0.008818	0.074580
Flow_Direction	-0.004331	1.000000	0.001414	0.009730	0.014564	0.012535	0.037115	0.025257	-0.034905
Vehicle_category	0.004000	0.001414	1.000000	0.011047	-0.026008	-0.024754	-0.009838	-0.044585	0.183555
Fix_Obst_st	0.004541	0.009730	0.011047	1.000000	-0.328615	-0.030390	0.007594	-0.019914	0.243668
Mov_Obst_st	0.001722	0.014564	-0.026008	-0.328615	1.000000	0.028057	0.009547	0.015129	-0.077708
Shock_pt	-0.000754	0.012535	-0.024754	-0.030390	0.028057	1.000000	0.099523	0.018520	-0.061628
Maneuver	0.003393	0.037115	-0.009838	0.007594	0.009547	0.099523	1.000000	0.004503	-0.036842
motor	0.008818	0.025257	-0.044585	-0.019914	0.015129	0.018520	0.004503	1.000000	-0.011269
No.of_occupants	0.074580	-0.034905	0.183555	0.243668	-0.077708	-0.061628	-0.036842	-0.011269	1.000000

Non-numeric columns: Vehicle_No -> should be numeric

Notices:Some variables start with -1, this value usually stands for "not known" or "not assigned". Some variables seem to have a lot of values (e.g. Vehicle Category. This should be recoded, but unfortunately it was not done).

Characteristics:

```
<class 'pandas.core.frame.DataFrame'>
Index: 58713 entries, 201900000001 to 201900058840
Data columns (total 15 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Day           58713 non-null  int32
 1   Month         58713 non-null  int32
 2   Year          58713 non-null  int32
 3   Time          58713 non-null  float64
 4   Light         58713 non-null  float64
 5   Department    58713 non-null  object
 6   Municipality  58713 non-null  object
 7   Location      58713 non-null  float64
 8   int           58713 non-null  float64
 9   atm           58713 non-null  float64
 10  col           58713 non-null  float64
 11  adr           58286 non-null  object
 12  lat           58713 non-null  float64
 13  long          58713 non-null  float64
 14  Date          58713 non-null  object
dtypes: float64(8), int32(3), object(4)
memory usage: 8.5+ MB
```

	Day	Month	Year	Time	Light	Location	int	atm	col	lat	lon
unt	58713.000000	58713.000000	58713.0	58713.000000	58713.000000	58713.000000	58713.000000	58713.000000	58713.000000	5.871300e+04	5.871300e+0
ean	15.685964	6.688485	2019.0	0.576343	1.945021	1.645990	2.025480	1.617052	4.164717	4.458634e+08	2.864451e+0
std	8.720109	3.387156	0.0	0.230809	1.504425	0.478216	1.997358	1.677022	1.953664	1.182908e+08	1.871112e+0
nin	1.000000	1.000000	2019.0	0.000000	1.000000	1.000000	1.000000	-1.000000	-1.000000	-6.142566e+08	-1.781574e+0
5%	8.000000	4.000000	2019.0	0.402778	1.000000	1.000000	1.000000	1.000000	3.000000	4.478586e+08	1.410280e+0
0%	16.000000	7.000000	2019.0	0.611111	1.000000	2.000000	1.000000	1.000000	3.000000	4.782554e+08	2.398705e+0
5%	23.000000	10.000000	2019.0	0.756944	3.000000	2.000000	2.000000	1.000000	6.000000	4.885652e+08	4.829060e+0
ax	31.000000	12.000000	2019.0	0.999306	5.000000	2.000000	9.000000	9.000000	7.000000	6.356807e+08	1.740234e+0

Non-numeric columns:

```
characteristics["Department"].nunique()
```

```
107
```

```
characteristics["Municipality"].nunique()
```

```
11421
```

```
characteristics["adr"].nunique()
```

```
31934
```

Correlations: no correlations worth mentioning

Anomalies:

- All data are from 2019.
- Time is coded between 0 and 1.
- Non-numeric columns "Municipality" and "adr" contain many different values.

Places:

```
<class 'pandas.core.frame.DataFrame'>
Index: 58840 entries, 201900000001 to 201900058840
Data columns (total 17 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   road_cat                   58814 non-null  float64
 1   route_num                  55879 non-null  object
 2   road_index1                48068 non-null  float64
 3   road_index2                4167 non-null   object
 4   traffic_regime             55669 non-null  float64
 5   traffic_lanes              58164 non-null  float64
 6   reserved_lane              58145 non-null  float64
 7   prof                       58800 non-null  float64
 8   upstream_terminal          51517 non-null  float64
 9   distance_to_upstream       51215 non-null  float64
 10  plan                       58803 non-null  float64
 11  central_reservation_width  211 non-null    float64
 12  roadwidth_vehicles         372 non-null    float64
 13  surf                       58793 non-null  float64
 14  infra                      58769 non-null  float64
 15  situ                       58698 non-null  float64
 16  speed_limit                57933 non-null  float64
dtypes: float64(15), object(2)
memory usage: 10.1+ MB
```

	road_cat	road_index1	traffic_regime	traffic_lanes	reserved_lane	prof	upstream_terminal	distance_to_upstream	plan	centr
count	58814.000000	48068.000000	55669.000000	58164.000000	58145.000000	58800.000000	51517.000000	51215.000000	58803.000000	
mean	3.347740	0.008758	1.980636	2.443247	0.149058	1.244184	19.138944	189.394709	1.305682	
std	1.250984	0.137390	0.605018	1.401903	0.574300	0.558611	111.939088	314.454298	0.679804	
min	1.000000	0.000000	1.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	
25%	3.000000	0.000000	2.000000	2.000000	0.000000	1.000000	0.000000	0.000000	1.000000	
50%	3.000000	0.000000	2.000000	2.000000	0.000000	1.000000	0.000000	0.000000	1.000000	
75%	4.000000	0.000000	2.000000	3.000000	0.000000	1.000000	13.000000	336.000000	1.000000	
max	9.000000	3.000000	4.000000	12.000000	3.000000	4.000000	8246.000000	7000.000000	4.000000	

upstream_terminal	distance_to_upstream	plan	central_reservation_width	roadwidth_vehicles	surf	infra	situ	speed_limit
51517.000000	51215.000000	58803.000000	211.000000	372.000000	58793.000000	58769.000000	58698.000000	57933.000000
19.138944	189.394709	1.305682	4.684360	42.054301	1.279251	0.881740	1.408344	60.854297
111.939088	314.454298	0.679804	16.141514	64.433320	0.855923	2.248355	1.241448	22.927331
0.000000	0.000000	1.000000	0.000000	2.400000	1.000000	0.000000	1.000000	0.000000
0.000000	0.000000	1.000000	0.000000	6.000000	1.000000	0.000000	1.000000	50.000000
0.000000	0.000000	1.000000	0.000000	14.000000	1.000000	0.000000	1.000000	50.000000
13.000000	336.000000	1.000000	2.000000	65.000000	1.000000	0.000000	1.000000	80.000000
8246.000000	7000.000000	4.000000	180.000000	730.000000	9.000000	9.000000	8.000000	800.000000

Non-numeric columns:

```
places["route_num"].nunique()
```

```
14327
```

```
places["road_index2"].nunique()
```

```
35
```

Correlations: Speed_limit---road_cat: -0.550833 -> is as expected.

Anomalies:

- Many zero values in the columns upstream_terminal and distance_to_upstream.
- In speed_limit: values up to 800.
- Many missing values in the columns central_reservation_width and roadwidth_vehicles.
- Many missing values in road_index2
- Many different values in route_num

Drivers:

```
<class 'pandas.core.frame.DataFrame'>
Index: 132977 entries, 138 306 524 to 137 982 130
Data columns (total 14 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   ID_Acc              132977 non-null  int64
 1   Vehicle_No          132977 non-null  object
 2   place               132977 non-null  int64
 3   User_category       132977 non-null  int64
 4   Severity            132977 non-null  int64
 5   Gender_driver       132977 non-null  int64
 6   DOB_driver          132977 non-null  int64
 7   Reason_for_travel   132977 non-null  int64
 8   Safety_equip_1      132977 non-null  int64
 9   Safety_equip_2      132977 non-null  int64
 10  Safety_equip_3      132977 non-null  int64
 11  Ped_loc             132977 non-null  int64
 12  Ped_act             132977 non-null  object
 13  Ped_acc             132977 non-null  int64
dtypes: int64(12), object(2)
memory usage: 15.2+ MB
```

	ID_Acc	place	User_category	Severity	Gender_driver	DOB_driver	Reason_for_travel	Safety_equip_1	Safety_equip_2	Safety
count	1.329770e+05	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000	13297
mean	2.019000e+11	2.184754	1.352587	2.542635	1.320304	1980.079600	3.224279	2.179790	1.454282	-
std	1.699444e+04	2.695724	0.630635	1.371815	0.466595	19.164625	2.667380	2.474627	3.284485	
min	2.019000e+11	1.000000	1.000000	1.000000	1.000000	1900.000000	-1.000000	-1.000000	-1.000000	-
25%	2.019000e+11	1.000000	1.000000	1.000000	1.000000	1967.000000	0.000000	1.000000	-1.000000	-
50%	2.019000e+11	1.000000	1.000000	3.000000	1.000000	1983.000000	4.000000	1.000000	0.000000	-
75%	2.019000e+11	2.000000	2.000000	4.000000	2.000000	1995.000000	5.000000	2.000000	3.000000	-
max	2.019001e+11	10.000000	3.000000	4.000000	2.000000	2019.000000	9.000000	9.000000	9.000000	

lser_category	Severity	Gender_driver	DOB_driver	Reason_for_travel	Safety_equip_1	Safety_equip_2	Safety_equip_3	Ped_loc	Ped_acc
32977.000000	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000	132977.000000
1.352587	2.542635	1.320304	1980.079600	3.224279	2.179790	1.454282	-0.913361	-0.109267	-0.810306
0.630635	1.371815	0.466595	19.164625	2.667380	2.474627	3.284485	0.892062	1.247776	0.648306
1.000000	1.000000	1.000000	1900.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
1.000000	1.000000	1.000000	1967.000000	0.000000	1.000000	-1.000000	-1.000000	-1.000000	-1.000000
1.000000	3.000000	1.000000	1983.000000	4.000000	1.000000	0.000000	-1.000000	0.000000	-1.000000
2.000000	4.000000	2.000000	1995.000000	5.000000	2.000000	3.000000	-1.000000	0.000000	-1.000000
3.000000	4.000000	2.000000	2019.000000	9.000000	9.000000	9.000000	9.000000	9.000000	3.000000

Non-numeric columns: Vehicle_No, Ped_act -> should be numeric.

Correlations:

- User_Category---place: 0.897206-> driver is always sitting on the driver's seat
- User_Category---Ped_loc: 0.651889 -> pedestrian is on the sidewalk
- Ped_acc---Ped_loc: 0.779813-> next to the road are more often groups

Anomalies:

- Some variables start with -1.
- Oldest driver is 119 years old.

Missing Values:

	features	missing_rate
0	Day	0.002
1	Month	0.002
2	Year	0.002
3	Time	0.002
4	Light	0.002
5	Department	0.002
6	Municipality	0.002
7	Location	0.002
8	int	0.002
9	atm	0.002
10	col	0.002
11	adr	0.0097
12	lat	0.002
13	long	0.002
14	Date	0.002
15	road_cat	0.002
16	route_num	0.0464
17	road_index1	0.1863
18	road_index2	0.927
19	traffic_regime	0.0523
20	traffic_lanes	0.0125
21	reserved_lane	0.0127
22	prof	0.0022
23	upstream_terminal	0.1204
24	distance_to_upstream	0.1255
25	plan	0.0021
26	central_reservation_width	0.9965
27	roadwidth_vehicles	0.9938
28	surf	0.0023
29	infra	0.0027
30	situ	0.0037
31	speed_limit	0.0162
32	Vehicle_No	0.0
33	place	0.0
34	User_category	0.0
35	Severity	0.0
36	Gender_driver	0.0
37	DOB_driver	0.0
38	Reason_for_travel	0.0
39	Safety_equip_1	0.0
40	Safety_equip_2	0.0
41	Safety_equip_3	0.0
42	Ped_loc	0.0
43	Ped_act	0.0
44	Ped_acc	0.0
45	Flow_Direction	0.0
46	Vehicle_category	0.0
47	Fix_Obst_st	0.0
48	Mov_Obst_st	0.0
49	Shock_pt	0.0
50	Maneuver	0.0
51	motor	0.0
52	No.of_occupants	0.9859
53	DayOfWeek	0.002

Anomalies: Some columns have a lot of missing values (e.g. No.of_occupants). More detailed information can be found in the Data Preparation section.

3.3 Summary of Key findings

- Some columns are strongly affected by missing values.
- The coding of the categorical variables is inconsistent. Some start at 0, some at -1. The coding is not always ascending by 1.
- Some variables have very many expressions, but cannot be grouped together.
- Overall little correlation of the columns (in the respective data sets).

3.4 Data contents

The type of road (correlated with speed_limit), the user_category (driver, passenger or pedestrian) should be decisive for the severity of the consequences of the accident. Weather conditions should also play a role. It can also be assumed that missing safety equipment aggravates the consequences of the accident. If rollovers happen during the accident (shock points: multiple) one can expect a serious accident. In car accidents, the consequences should be less severe than in motorcycle accidents, because the car offers more protection to the driver. If the road makes a curve, the consequences of the accident could be worse than on a straight road.

Although more accidents happen in large cities (e.g. Department 75 = Paris), it may be that these are mainly minor accidents. We can only speculate about many columns:

1. Do more serious accidents happen on the way to work or during work-related trips?
2. Are the consequences of accidents worse for internal combustion or electric cars?
3. Do worse accidents happen depending on which direction a road is traveled?
4. Which features are most likely to affect the outcome?
5. Are there different clusters in the data that lead to the same target column?

4.Data Preparation

First, the characteristics and places files are merged using the common ID_ACC (accident number) column.

From the columns for year, day and month a column date was generated.

After that the files drivers and vehicles are joined by the common column ID_veh (vehicle number).

Afterwards all files can be joined together over the common column ID_Acc.

This creates a large dataset with 132977 rows and 53 columns. From the column of the date a new column weekday was calculated.

Missing values in the columns were counted. Many columns contain missing values, some even over 90%. It should be noted that missing values were also coded by -1 or 0 in the data set. For example, in addition to 12% missing values, the distance_to_upstream column mainly contains the value 0. Due to too many missing values, the columns "road_index1", "road_index2", "upstream_terminal", "distance_to_upstream", "central_reservation_width", "roadwidth_vehicles", "No.of_occupants" and "Safety_equip_3" were removed.

Other columns are no longer needed for analysis. These were Date and Year

Additionally, rows were removed with that had missing values in certain columns. This affected the columns Day, prof, plan, surf, adr, route_num and speed_limit. In these columns the missing values cannot be replaced in a meaningful way. For example, it makes no sense to replace the day or the number of the road (route_num) with the median. Also central characteristics of the accident like

the state of the road surface (surf), or the course of the road (plan) should not be replaced by the median, because they are categorical. Furthermore, with these important determinants of the accident, a bias of the result is to be feared by a bad substitution. Since the data set is large, we omit these rows.

In other columns, replacements were made to avoid losing too many records:

Here, missing values were replaced by the mode that was also the most normal or least marked value.

In the column traffic_regime missing values were coded with the value for road with two-way traffic (as opposed to one-way road or separated dirt road).

In the column "traffic_lanes" the value 2 (for two lanes) was chosen as replacement.

In the "reserved_lane" column, the value 0 (for no reserved lane as opposed to bike lane or bike path) was chosen as the replacement.

In the "infra" column, the value 0 (for no infrastructure as opposed to bridge or tunnel) was chosen as the replacement.

In the column "situ", the value 0 (for accident on road as opposed to accident on bike lane or on sidewalk) was chosen as replacement.

In the "speed_limit" column, some values were combined: Speed limits above 300 were coded as 300, thus indicating that there was no speed limit. Speed limit below 7 were coded as 7.

In the column "ped_act" (pedestrian action) there were not only numerical values , but also letters. These were assigned numeric values. The reason for this is probably a change in the coding guidelines.

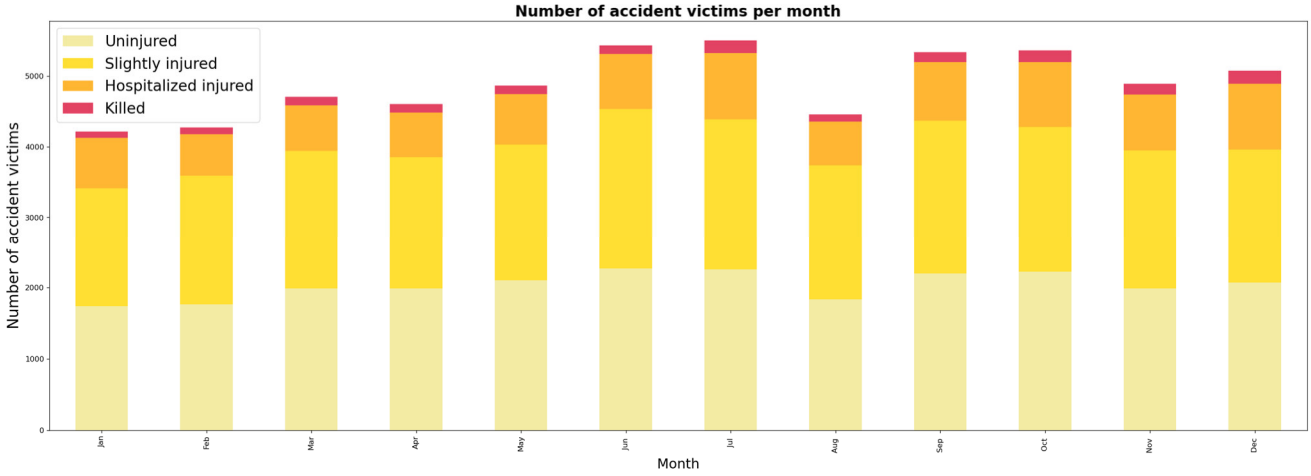
To increase the comparability of the columns, the coding was changed in some columns to avoid negative values. The value -1 in columns with categorical columns stands for "not assigned" in these columns. In addition, in these columns there were the expressions "unknown" or "not ascertainable". The values -1 were therefore recoded to the these other expressions. This affected the columns Reason_for_travel (reason for travel), Safety_equip1, Safety_equip_2 (safety equipment), "Ped_loc" (location of pedestrian), "Flow_Direction" (direction in which the road was traveled), "Maneuver" (action of the driver before the accident), and "motor" (motor type).

Columns with non-numeric content were numerically encoded using the LabelEncoder from sklearn.

After cleaning, we get a dataset with 124382 rows and 44 columns.

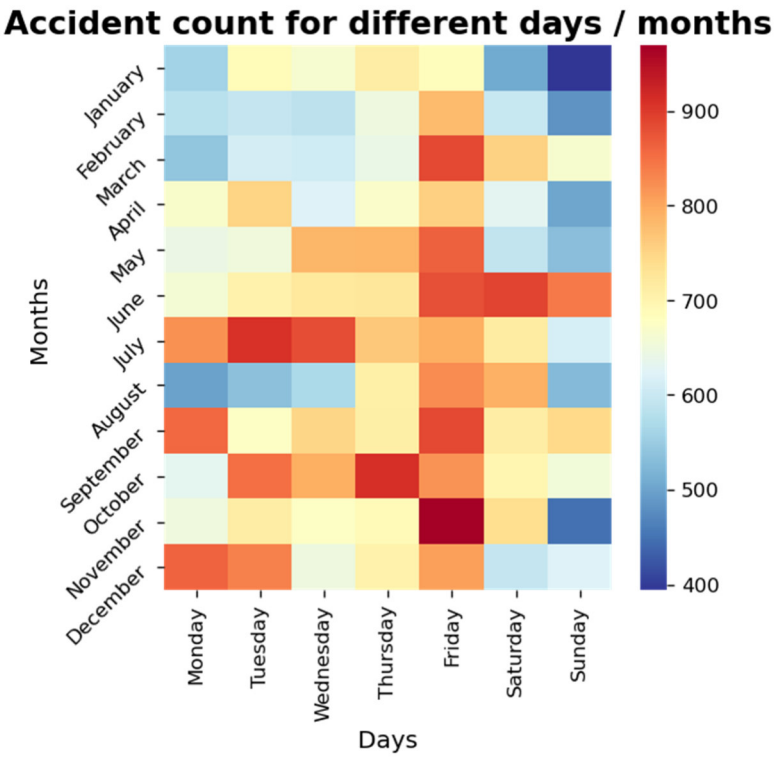
5. Exploratory Data Analysis

Is there a trend in number of accidents according to Monthly seasonlaity?



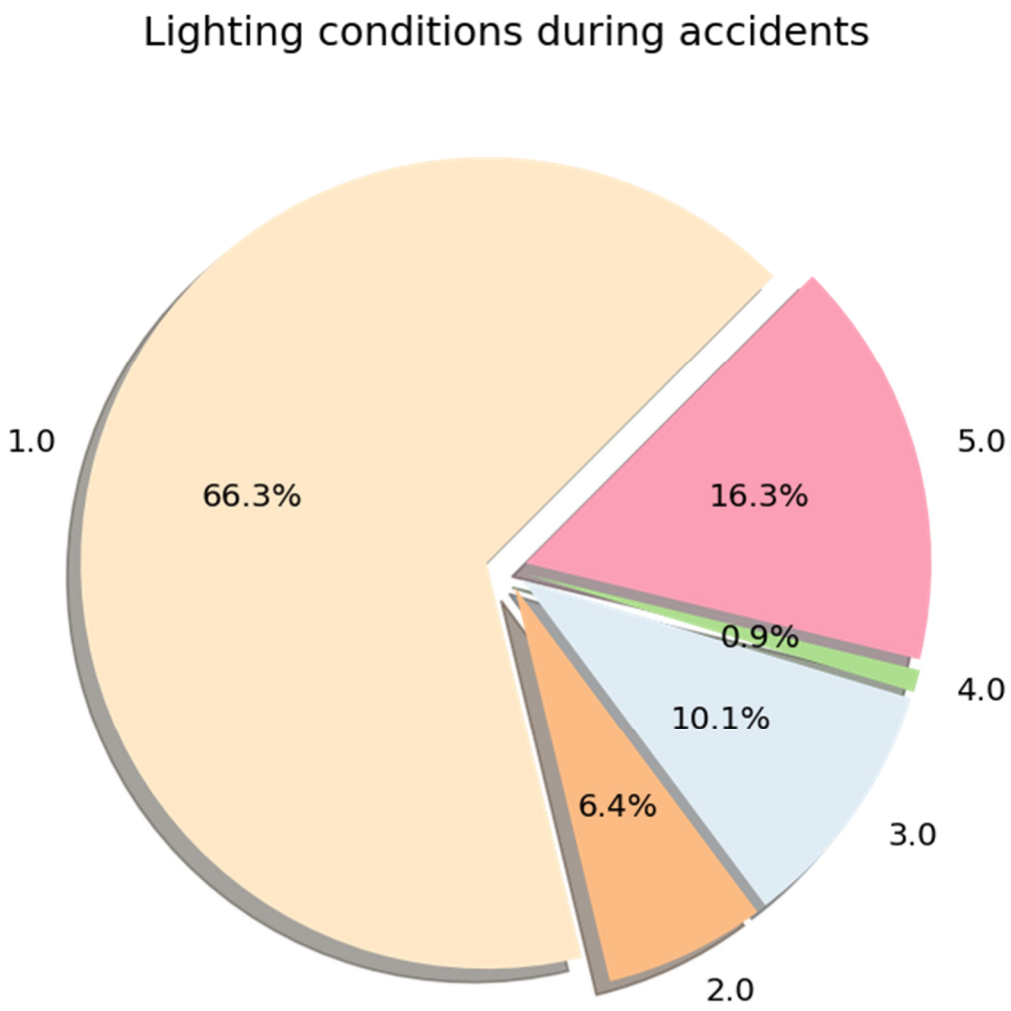
- A sharp rise is observed in the months of June, July, Septemeber and October
- There is steep reduction in no.of accidents for the month of January, February & August..
- July 2019 reported the highest number of accidents
- August 2019 reported lowest number of accidents

Does the Day of the week influences the series of accidents for each Month?



- Friday of October reported highest number of accidents
- Sunday of January reported less number of accidents
- Friday influences highly in the number of accidesnts throughout the year except for the month of January.

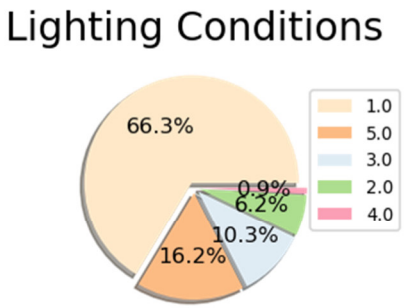
Does Lighting Conditions have any impact on Severity of accidents?

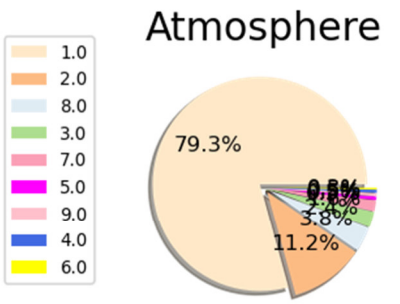


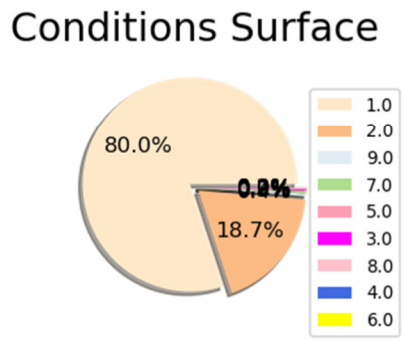
Light:
1. Full day
2. Twilight or dawn
3. Night without public lighting
4. Night with public lighting not on
5. Night with public lighting on

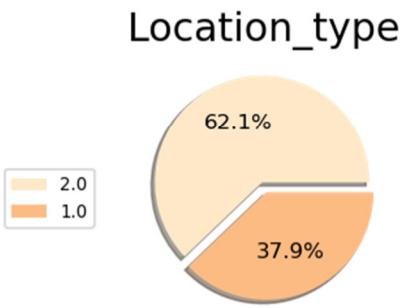
- Majority of accidents is reported under good lighting conditions (either 'Full Day' or at 'Night with public lighting on')

Good Lighting conditions reports to have high number of accidents, then would other factors play an important role with Lighting?





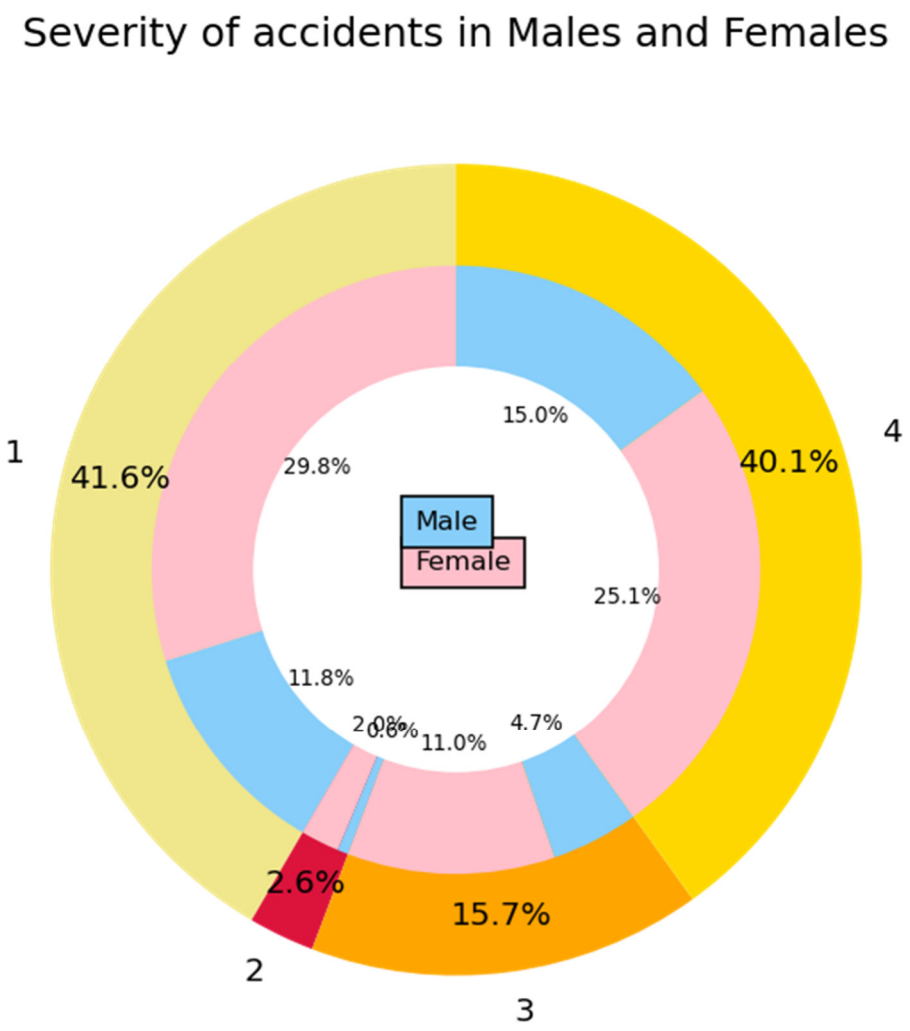




Atmospheric conditions:	Location:	Light: lighting conditions in which the accident occurred:	Surface condition:
1. Normal	1. Outside Urban agglomeration	1. Full day	1. Normal
2. Light rain	2. In built-up areas	2. Twilight or dawn	2. Wet
3. Heavy rain		3. Night without public lighting	3. Puddles
4. Snow. hail		4. Night with public lighting not on	4. Flooded
5. Fog. smoke		5. Night with public lighting on	5. Snowy
6. Strong wind. storm			6. Mud
7. Dazzling weather			7. Icy
8. Cloudy weather			8. Fat. oil
9. Other			9. Other

- Lighting conditions as already observed reported accidents under good lighting conditions
- Atmosphere didn't play an important role in the increase in number of accidents
- Similarly, normal surface conditions reported the highest number of accidents of 80% ; and hence the surface conditions also didn't influence the numbers.
- As expected the in-built areas reported 62.1% of accidents compared to 37.9% outside urban areas

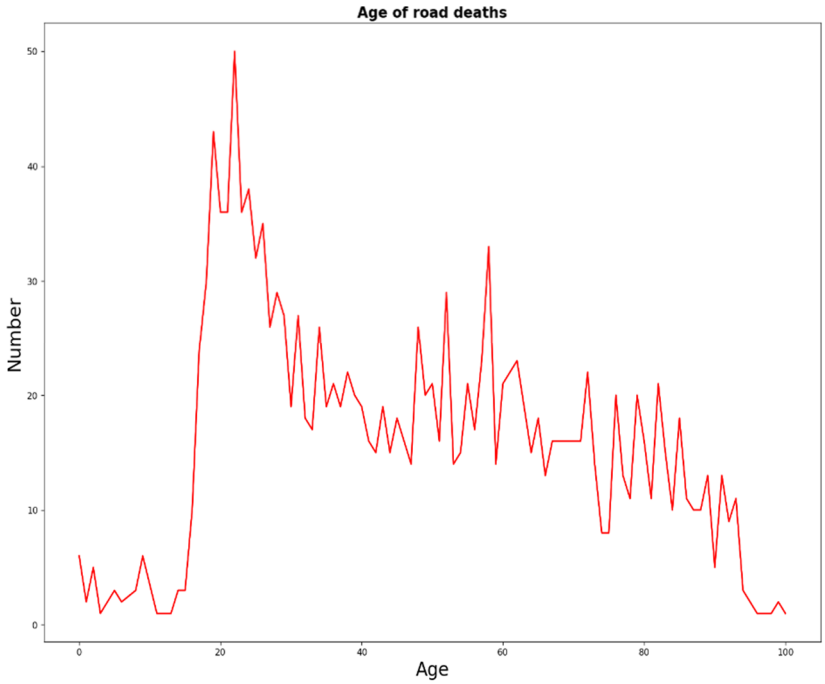
Did Gender feature observed any bias in the severity of accident types?



Severity
1. Unharmed
2. Killed
3. Injured Hospitalized
4. slight Injury

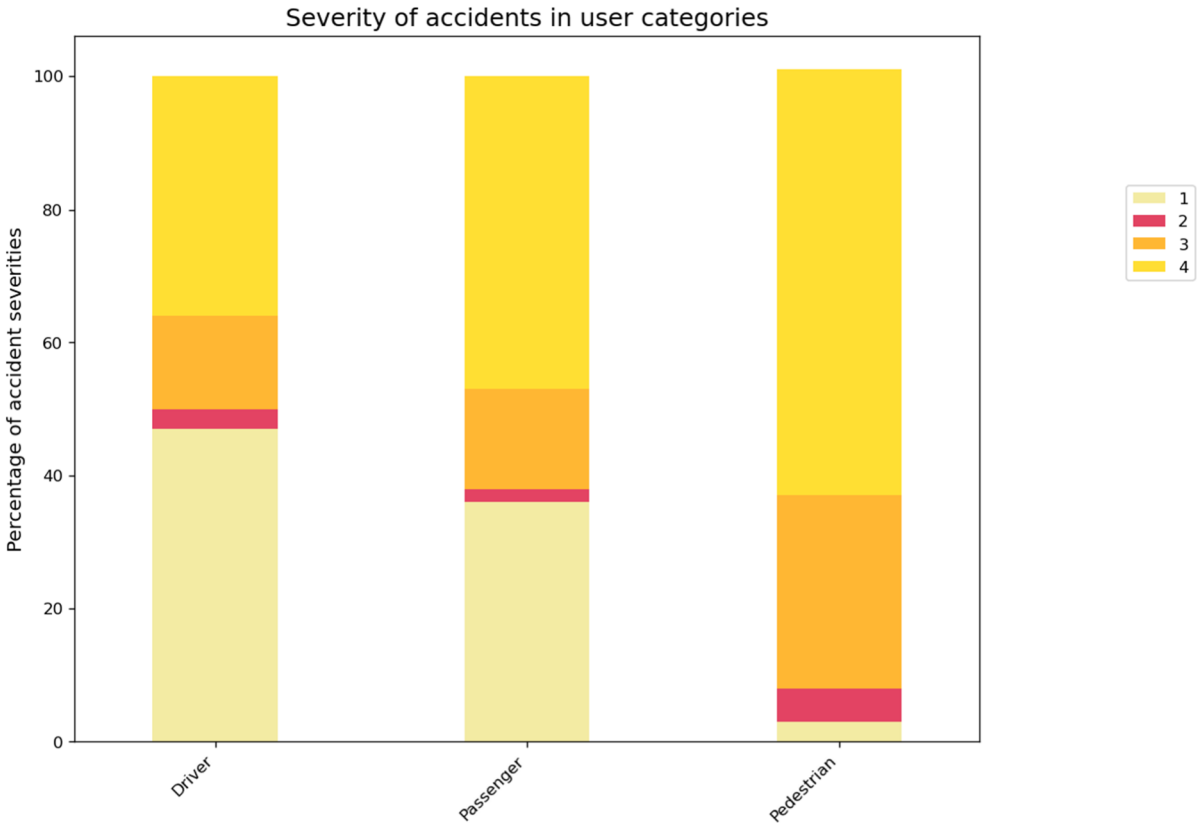
- Majority of users come out unscathed (41.6%) or incur only light injuries (40.1%) during an accident.
- The proprotion of Males vs Females under each severity category, are roughly similar

Distribution of Drivers involved in accidents by Age.



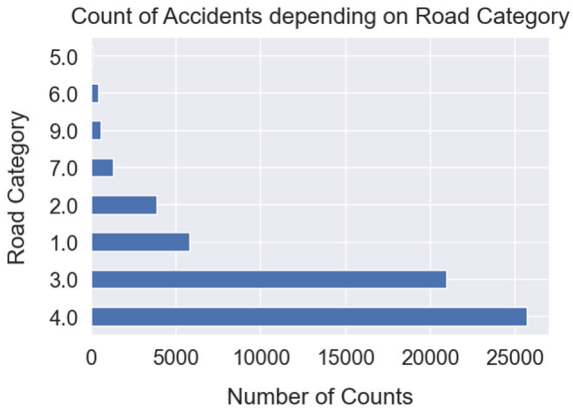
- Maximum Drivers with age range 16-26 have reported death in road accidents

Which road users are highly prone to severity of road accidents?



- As expected, Pedestrians are more likely to incur injuries or get hospitalized during accidents as compared to users who are Drivers or Passengers.

Are there high accidents in Highway?

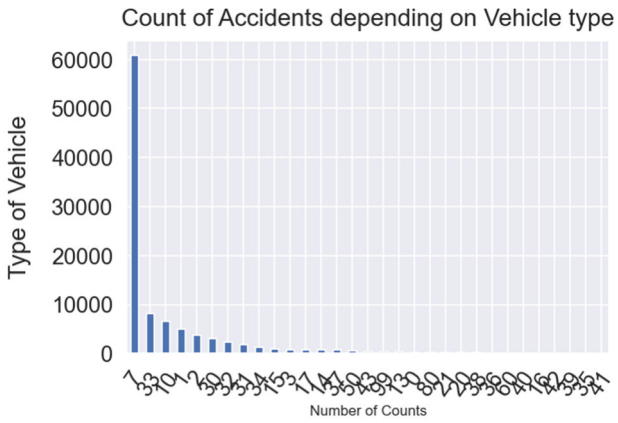


Road Category leading to road traffic accidents

- 4. Communal roads
- 3. Departmental road
- 1. Highway
- 2. National road
- 7. Urban metropolis roads

Highway road category did not result into highest numbers

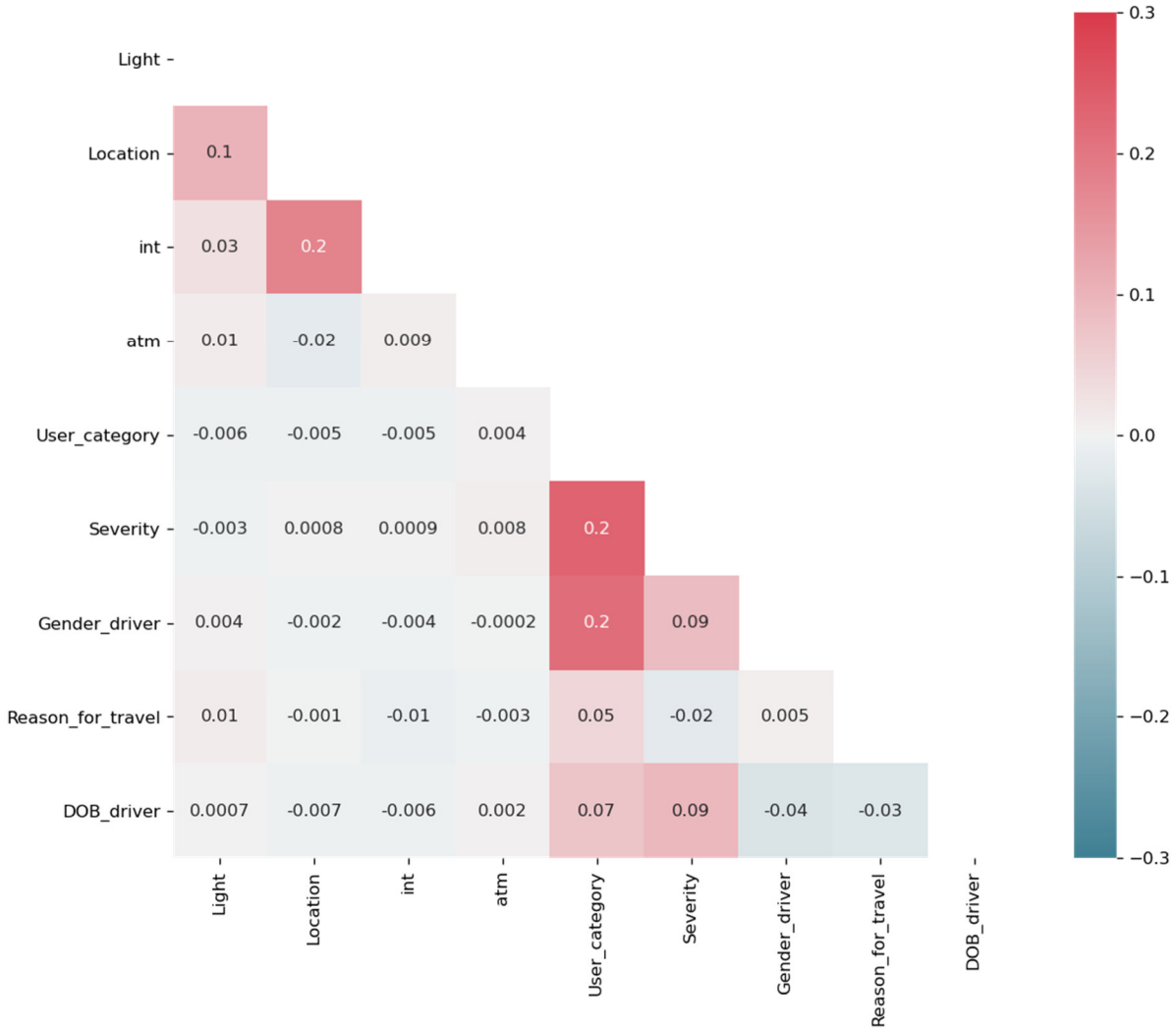
Which type of Vehicle has great severity in road accident?



Road Category leading to road traffic accidents

- 7. Car
- 33. Tram
- 10. Light vehicles + trailer
- 1. Bicycle
- 2. Moped
- 30. Bus

Is there any correlation with other characteristics of road users with severity reported?



- User_category & Gender shows positive corelation to the severity of accidents
- Intersection and location also showed high correlation
- Light had no correlation with the severity of accidents

6. Application of ML Algorithms

6.1 PCA

The PCA on the whole dataset showed that almost all features are equally important, for this reason and the size of the dataset we first consider only "fast" algorithms, like Naive Bayes.

Due to already very poor scores, we also tried KNN and RandomForest. The respective gridSeaches ran for 6h and then returned scores of about 0.6 on the test set with the best parameters.

In order to be able to try more algorithms and not have to wait 30-60min for each fit, we restricted our working hypothesis somewhat. In the following, we no longer consider the severity of all accidents in France, but only each one on the highway. From 130000 lines, thus became about 12000. Moreover, PCA revealed that two features are responsible for 99% of the variance or information. Consequently, we considered only the PCA fit of the two features and only about 10% of the lines.

Most algorithms scale linearly with the number of features, i.e., from 46 before to 2 now reduces the computation time by 23 times.
Reducing the dataset to 10% of the original rows has a huge impact on computation time. Naive Bayes scales with O(n), the RandomForest with O(n log(n)) and the SVC even with $O(n^2)$.

SVC also takes an enormous amount of time because no multithreading is possible with this algorithm. A 2 hours lasting first GridSearch brought only minimal improvements in the score, so that we did not treat this Classifer further.

For the scores, the name is first, then the score on the training set, and then on the test set. Only the accuarcy score is considered.

```
In [86]: scores

Out[86]: [('GaussianNB()', 0.45838533166938655, 0.4618082618862042),
          ('KNeighborsClassifier()', 0.630795814533935, 0.4520654715510522),
          ('RandomForestClassifier()', 0.7253527887107613, 0.41095089633671084)]
```

Scores with stratified training set:

```
In [90]: scores

Out[90]: [('GaussianNB()', 0.4535854852644715, 0.4553780202650039),
          ('KNeighborsClassifier()', 0.6371316117884228, 0.45654715510522215),
          ('RandomForestClassifier()', 0.9998080061438034, 0.6718628215120811)]
```

No overfitting is present.

6.2 Refinement

Scores mean for a crossvalidation with 5 foldes (on the test set):

```
In [104]: cv_results

Out[104]: [('GaussianNB()', 0.4553777064103513),
           ('KNeighborsClassifier()', 0.4694076693410471),
           ('RandomForestClassifier()', 0.6533518964564934)]
```

The above scores seem to be correct.

7. Gridsearch

Description and estimation of the results

The image shows the scores of the classfiers with the best hyperparameters from GridSearch. The first row is KNN and the second is RandomForest. For Random Forest, we could see a slight improvement on the test set.

```
In [143]: gs_scores

Out[143]: [(0.9076509551694346, 0.6769290724863601),
           (0.5834693289814726, 0.4696024941543258)]
```

8. Use of an ensemble method

We use a voting classifier. First with the three classifiers mentioned above and hard-voting, then soft-voting. Then we tried adding the SVC to the three classifiers and finally we did a weighted soft-voting, since the RandomForestClassifier already gave the highest score. However, the RandomForrestClassifer always has a higher score than the ensemble.

9. Final Results

Letztendlich wurde ein score von 0.6718 durch den RandomForestClassifier erzielt.

In Anbetracht der Komplexität der Daten und der immer noch verbesserungswürdigen Kodierung der Features kann dies als gutes Ergebnis betrachtet werden. Der Datensatz ist zwar groß, jedoch ist nicht zu unterschätzen wie viele Werte eben doch für unbekannt, nicht zuzuordnen oder andere Kategorie standen. Nicht hilfreich scheinen Spalten mit sehr vielen Ausprägungen zu sein, da man über sie nicht abstrahieren kann. Es wäre darüber nachzudenken, ob man in gewissen Spalten Kategorien zusammenfassen kann (z.B. in Vehicle_Kategorie alle Zweiräder zusammenfassen) um weniger Unterschiede zu haben.