Road to Safety: Traffic Accident Analysis

Andrey Bessalov

ocean

Road to Safety: Traffic Accident Analysis

1. Challenge description
2. General Trends
    1. Yearly trends
    2. Yearly relative statistics per accident
3. Geographical Insights
    1. Correlation of countries population and the number of accidents
    2. TOP regions by the number of accidents
    3. TOP regions by the number of accidents per capita
4. Analysis of the most severe accidents
    1. Patterns of the most severe accidents
    2. Time patterns of the most severe accidents
    3. Road characteristics patterns of the most severe accidents
    4. Weather patterns of the most severe accidents
    5. Involved vehicles patterns of the most severe accidents
5. Temporal analysis and clustering
    1. Accident statistics by months
    2. Accident statistics by the day of week
    3. Accident statistics by the day hours
    4. Monthly clustering by the involved units distribution
    5. Yearly clustering by the involved units distribution
6. Prediction models development
    1. Modeling tasks description
    2. ARIMA modeling results

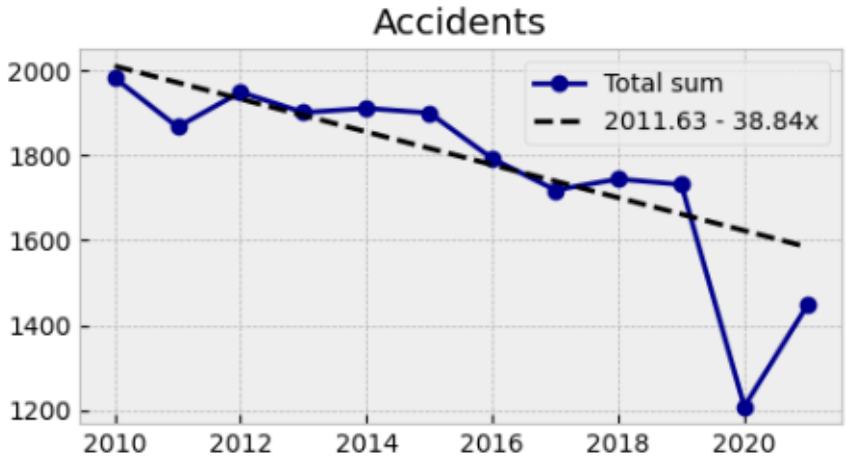
1. Challenge description

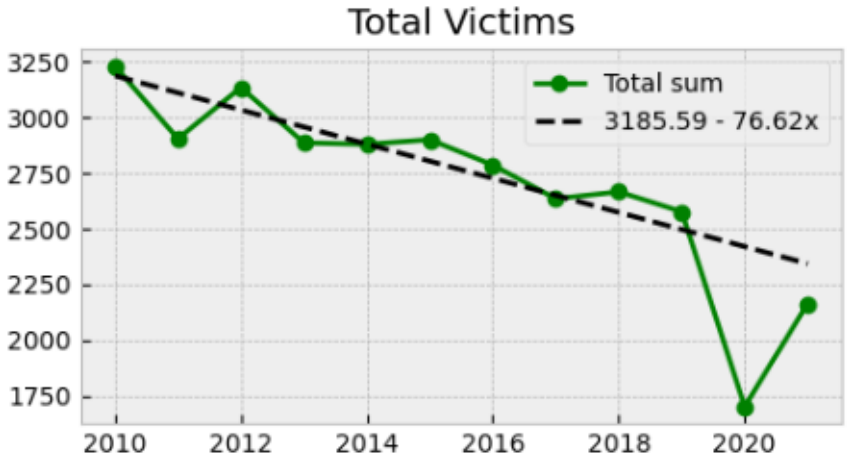
General Trends (5 points):
• What are the overall trends in traffic accidents, fatalities, and serious injuries in Catalonia from 2010-2021?
Accident Characteristics (5 points):
• What common characteristics (time of day, type of road, etc.) are observed in the most severe accidents?
Geographical Insights (5 points):
• Which municipalities or counties in Catalonia have the highest incidence of traffic accidents? How does this correlate with population density or road network characteristics?
Yearly Trends (5 points):
• How have traffic accident patterns (frequency, severity) changed yearly from 2010 to 2021?
Day and Time Patterns (5 points):
• On what days of the week and times of day do most accidents occur? Are there notable differences between weekdays and weekends?
Environmental Impact (5 points):
• How do different weather conditions affect the likelihood of accidents? Is there a correlation between visibility, road conditions, and accident severity?
Road and Traffic Features (5 points):
• What impact do road features (such as speed limits and road types) and traffic density have on the occurrence of accidents?
Vehicle Types and Accident Severity (5 points):
• Does the involvement of specific types of vehicles (like heavy trucks and motorcycles) correlate with more severe accidents?
Temporal Clustering (10 points):
• Are there specific periods (months, years) where accident patterns cluster significantly? What might be the causes for these clusters?
Time-Series Forecasting (20 points):
• Based on past trends, create a model to forecast the number of accidents, fatalities, or serious injuries for the upcoming year. Clearly describe the forecasting model you have developed. This should include the type of model, its structure, and any specific features or techniques it utilizes. Discuss the factors that influenced your decision, such as the model's accuracy, efficiency, suitability to the data characteristics, or its ability to handle the complexities of the dataset.
Report (30 points):
• Elaborate a comprehensive report incorporating analyses based on the previously outlined questions. The report should address each question, providing insights and findings derived from the data. Please ensure your report adheres to the format and guidelines specified in the 'Report Guidelines' section. It should include clear and well-structured sections corresponding to each analysis category, with relevant data visualizations and interpretations. The report should effectively communicate your analytical approach, key findings, and any conclusions or recommendations you draw from the data.

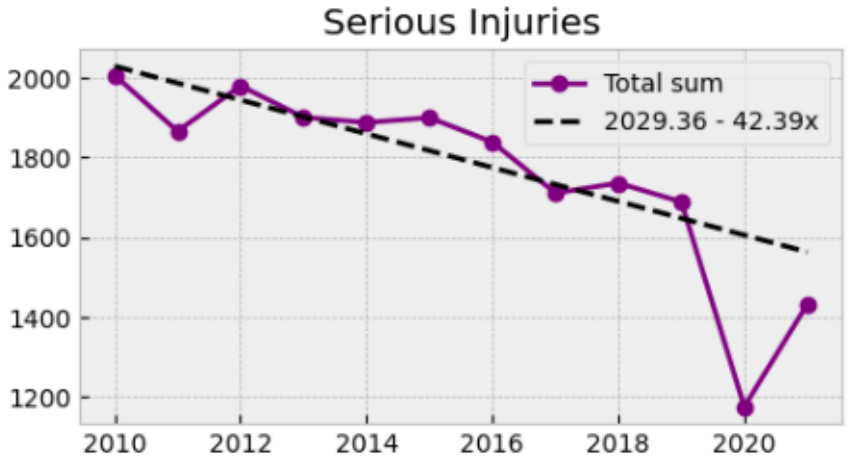
2.1. Yearly trends

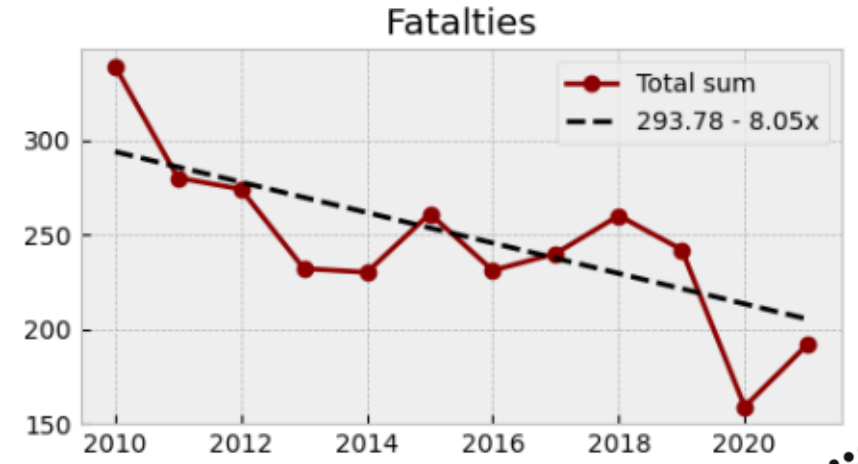
We are working with dataset of all accidents in Catalonia involving serious injury or death. The graphs below show the yearly number of those Accidents, Total Victims, Serious Injuries and Fatalities.

- We see that in 2020, due to COVID-19, there was a sharp drop in all statistics.
- When building trends, we did not take this year into account not to overestimate the rate of yearly decline.
- As we can see, there are a down-trending behavior for all indicators.









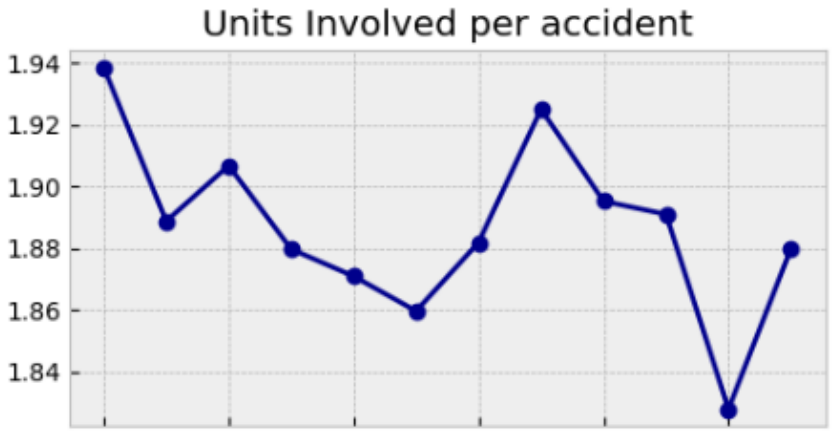
* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

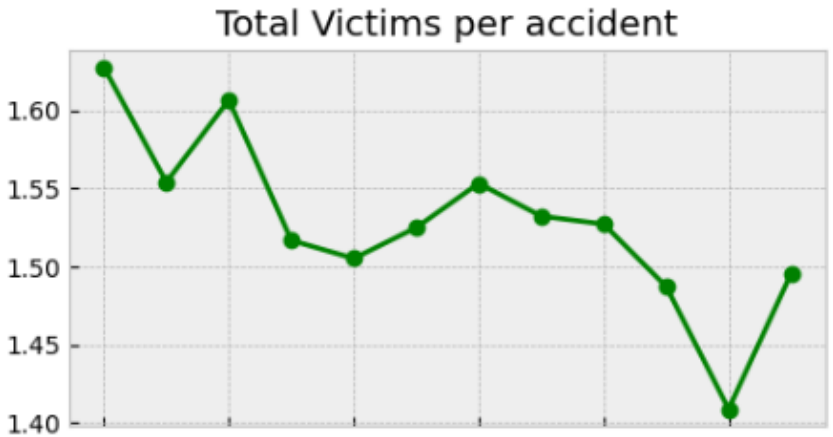
ocean

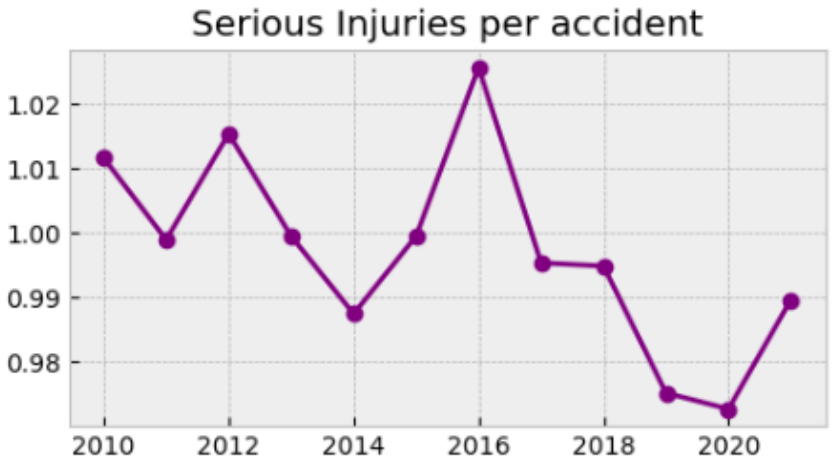
2.2. Yearly relative statistics per accident

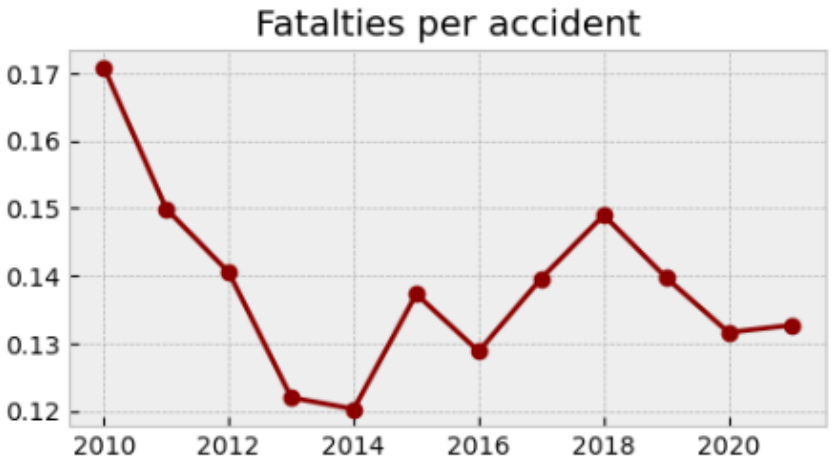
The graphs below show the yearly number of Units Involved, Total Victims, Serious Injuries and Fatalities per one accident.

- For all indicators there is a gradual decline, but not as significant as for absolute indicators.
- We see that in 2020, due to COVID-19, there was a sharp drop for Units Involved and Total Victims but not for Serious Injuries and Fatalities, meaning that the average number of victims and serious injuries per one accident does not highly correlated with the total number of accidents occurring on the roads.









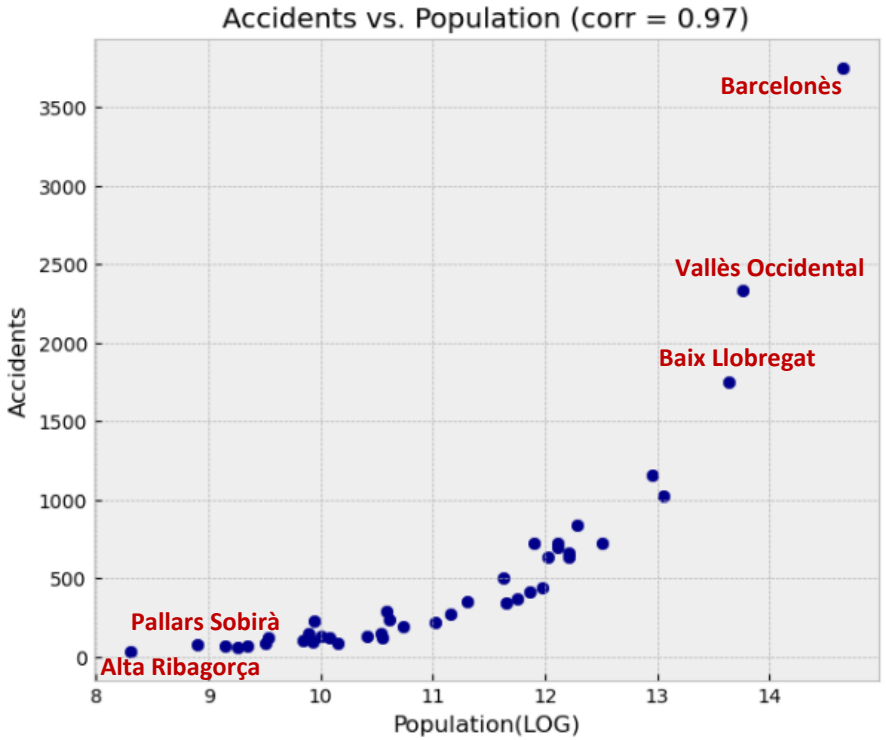
* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

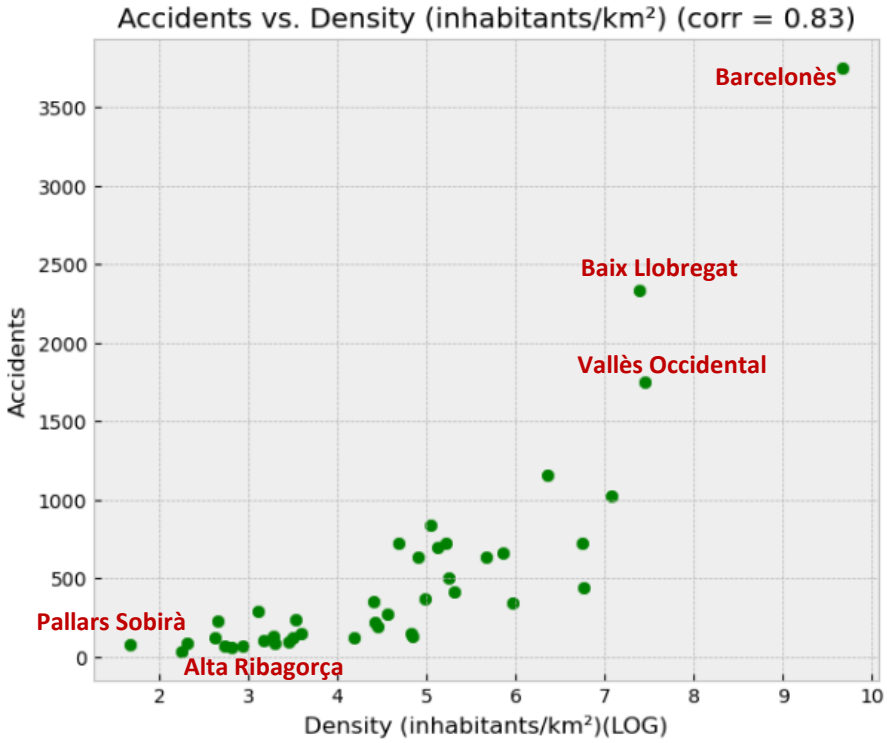
ocean

3.1. Correlation of countries population and the number of accidents

The graphs below show the dependencies of countries population and density with the number of accidents. On the x-axis placed a population indicator on the logarithmic scale, on the y-axis shown the number of accidents.
We downloaded population data from here: https://www.idescat.cat/indicadors/?id=aec&n=15227&lang=en

- Dependence between population and number of accidents is very high (Pearson's correlation coefficient is close to 1).
- Dependence between population density and number of accidents is also high.
- TOP regions with the highest population and the number of accidents: Barcelonès, Vallès Occidental, Baix Llobregat.
- BOTTOM regions with the lowest population and the number of accidents: Alta Ribagorça, Pallars Sobirà, Priorat.



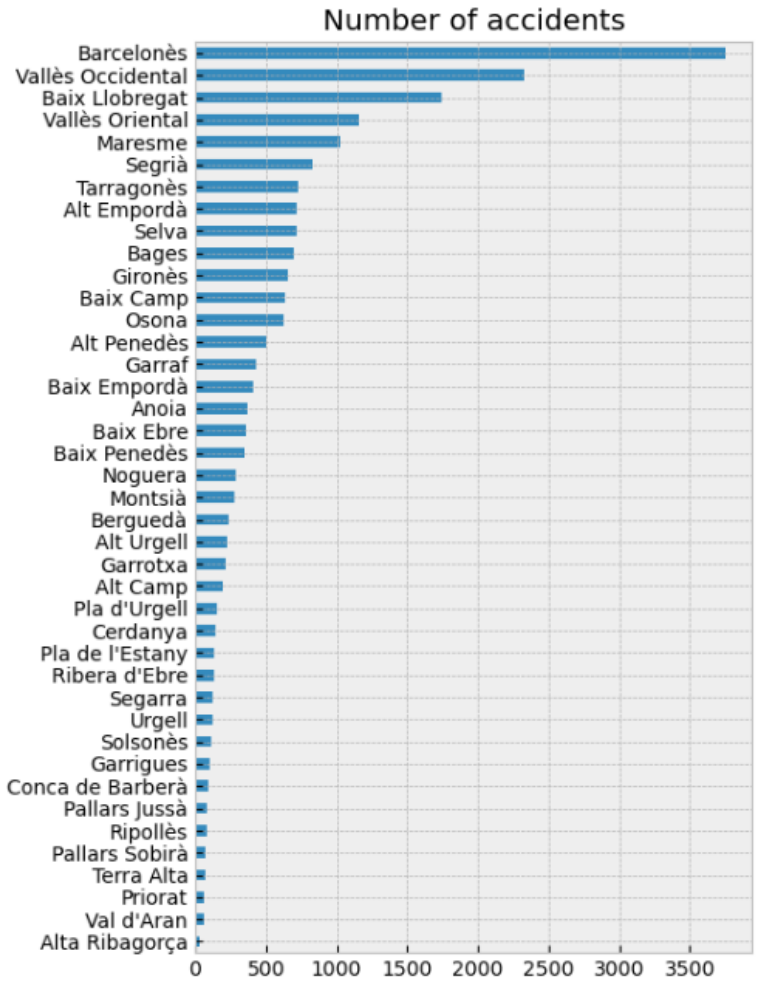


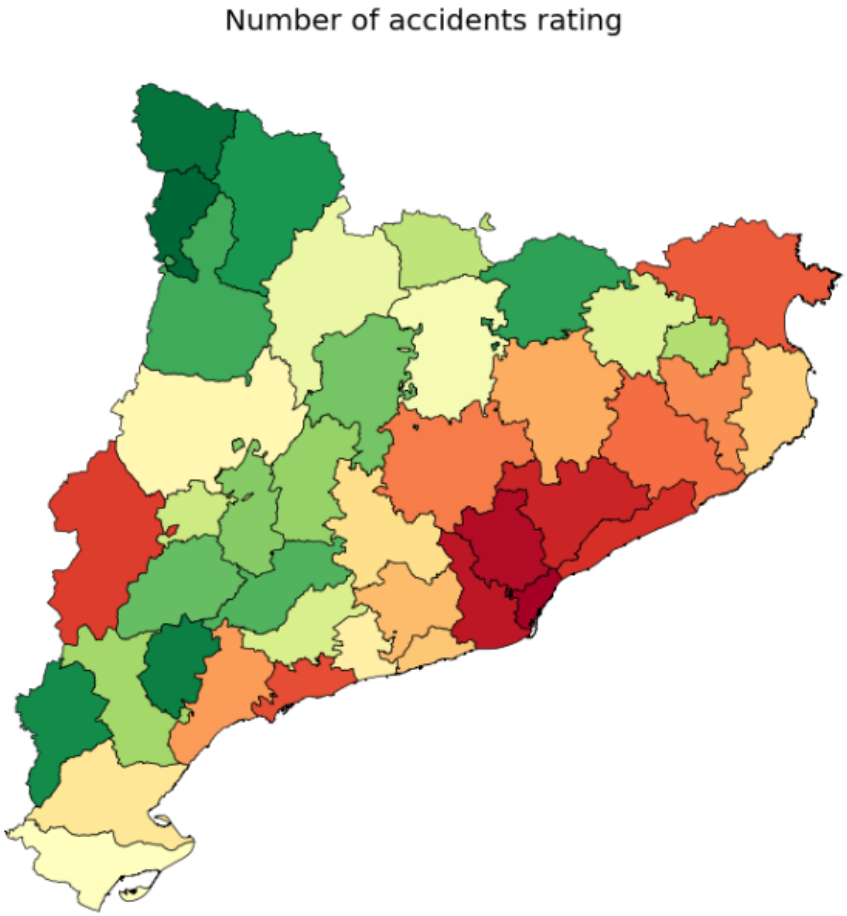
* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

ocean

3.2. TOP regions by the number of accidents

The pictures below show the number of accidents by regions as well as their ratings on the map (greener is better).
We see that the most accidents occur in the central regions near to Barcelona due to the fact that a large number of people live there.



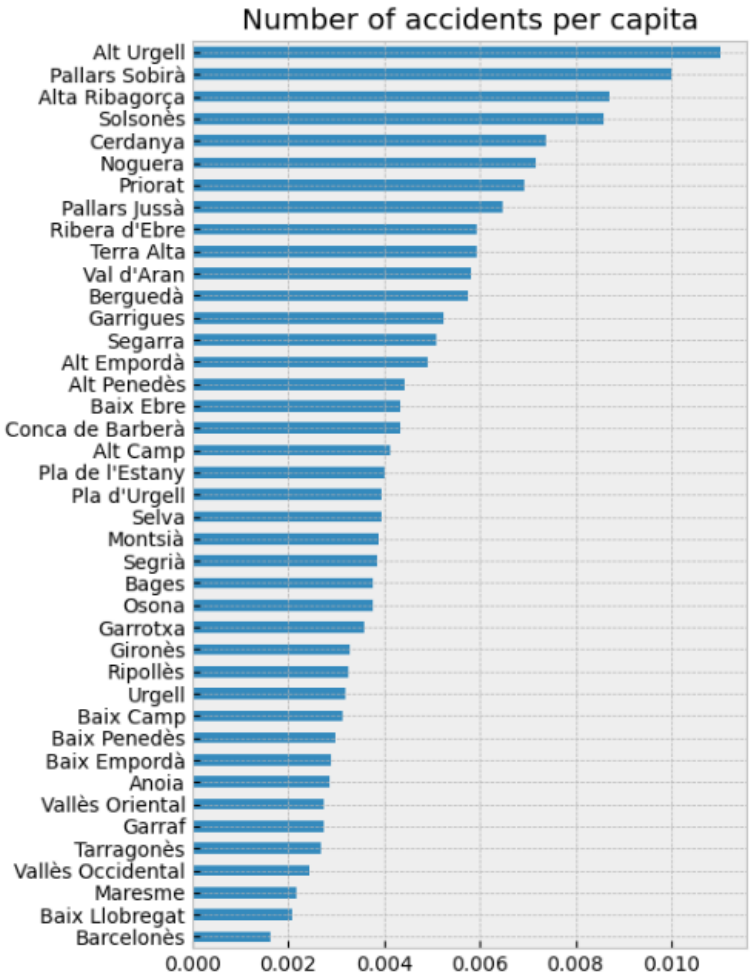


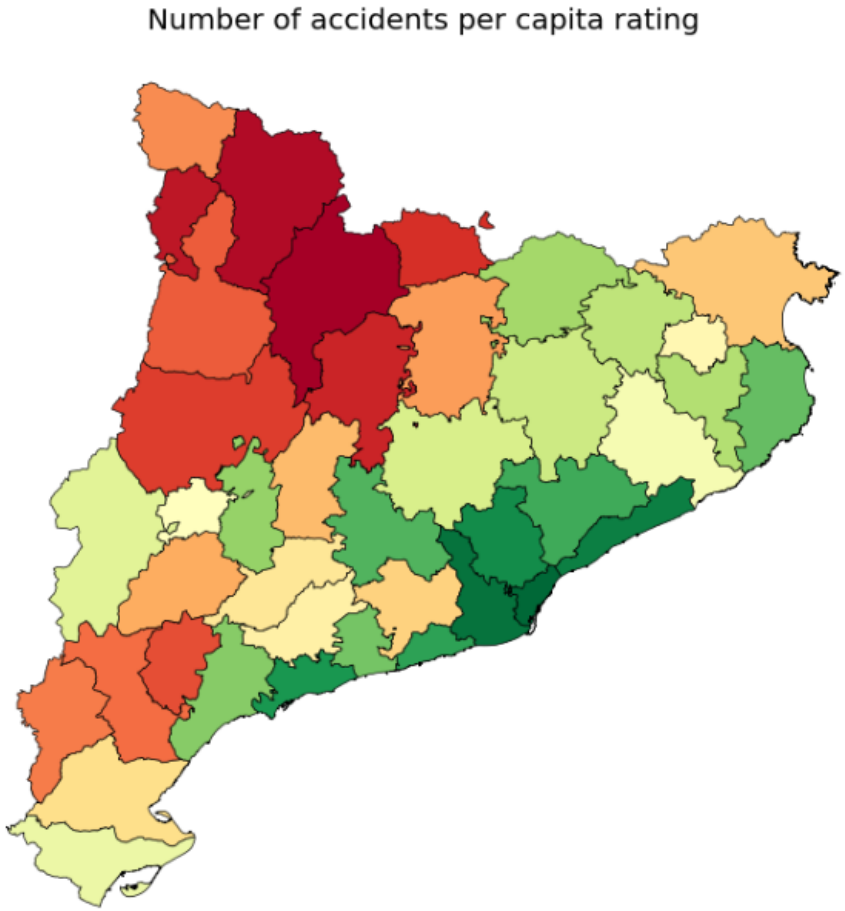
* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

ocean

3.3. TOP regions by the number of accidents per capita

The pictures below show the number of accidents per capita by regions as well as their ratings on the map (greener is better).
By comparing this statistics with the previous one we can see the opposite picture: Barcelona and their nearby regions have the lowest values of accidents per capita.





* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

ocean

4.1. Patterns of the most severe accidents

We calculated the following statistics to identify accidents popularity and patterns of the most severe accidents:

• **Feature** – feature describing the accident. We split features to the following areas: date/time, road description, weather conditions, different vehicles involved.
• **Value** – value of that feature. We took the values with the number of occurrences at least 100, which is about 0.5% of accidents.
• **Percent of accidents** – the percent of accidents with that feature value.
• **Probability of severe** – the conditional probability that the accident will be with fatalities.
• **Lift** – the ratio of the previous probability to the overall probability of accident with fatalities.
This probability calculated as: number of accidents with fatalities / number of accidents = 2673 / 21161 = 0.1263.

On the next slides we will show the most severe accident patterns by different features.

* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

4.2. Time patterns of the most severe accidents

- The most popular time when accidents occurring (by looking at the percent of accidents):
  - From May to July and October;
  - Friday, Saturday;
  - From 7:00 to 21:00 with the peak value at 18:00.
- The most severe accidents occurs in (by looking at the lift or probability of severity):
  - August, September, December;
  - Saturday and especially Sunday;
  - Night time: 22:00 – 6:00, with the peak value at 4:00.

Feature	Value	Percent of accidents	Probability of severe	Lift
Month	1	8%	12%	0,95
Month	2	7%	12%	0,95
Month	3	8%	13%	1,04
Month	4	8%	12%	0,92
Month	5	9%	11%	0,89
Month	6	9%	12%	0,93
Month	7	10%	13%	1,02
Month	8	8%	14%	1,11
Month	9	8%	14%	1,12
Month	10	9%	13%	1,01
Month	11	8%	12%	0,96
Month	12	8%	14%	1,10

Feature	Value	Percent of accidents	Probability of severe	Lift
Weekday	1	13%	12%	0,96
Weekday	2	14%	12%	0,95
Weekday	3	14%	12%	0,92
Weekday	4	14%	12%	0,97
Weekday	5	16%	12%	0,95
Weekday	6	15%	14%	1,08
Weekday	7	14%	15%	1,17

Feature	Value	Percent of accidents	Probability of severe	Lift
Hour	0	1%	14%	1,14
Hour	1	1%	21%	1,66
Hour	2	1%	20%	1,59
Hour	3	1%	26%	2,03
Hour	4	1%	31%	2,44
Hour	5	2%	20%	1,55
Hour	6	3%	22%	1,71
Hour	7	4%	14%	1,11
Hour	8	5%	12%	0,92
Hour	9	5%	12%	0,92
Hour	10	5%	11%	0,89
Hour	11	6%	10%	0,82
Hour	12	7%	10%	0,81
Hour	13	7%	10%	0,79
Hour	14	6%	12%	0,96
Hour	15	5%	11%	0,85
Hour	16	6%	12%	0,99
Hour	17	6%	11%	0,86
Hour	18	7%	11%	0,88
Hour	19	7%	12%	0,92
Hour	20	5%	12%	0,98
Hour	21	4%	13%	1,02
Hour	22	3%	16%	1,30
Hour	23	2%	17%	1,32

* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

ocean

4.3. Road characteristics patterns of the most severe accidents

The most severe accidents occurs with the following road features:

- On the highways (carretera);
- When speed limit > 60 km/h;
- When lighting conditions is bad;
- On the two-way roads (doble sentit).

Feature	Value	Percent of accidents	Probability of severe	Lift
Area	Carretera	45,5%	19%	1,51
Area	Zona urbana	54,5%	7%	0,58
Road Speed Limit	20	0,7%	6%	0,49
Road Speed Limit	30	4,7%	8%	0,67
Road Speed Limit	40	5,8%	9%	0,69
Road Speed Limit	50	5,6%	12%	0,96
Road Speed Limit	60	3,3%	17%	1,38
Road Speed Limit	70	1,7%	22%	1,74
Road Speed Limit	80	4,6%	21%	1,68
Road Speed Limit	90	1,1%	19%	1,52
Road Speed Limit	100	64,5%	15%	1,19
Road Speed Limit	120	0,5%	23%	1,81
Road Speed Limit	999	7,2%	2%	0,14
Lighting Conditions	Alba o capvespre	4,7%	16%	1,25
Lighting Conditions	De dia, dia clar	69,1%	11%	0,87
Lighting Conditions	De dia, dia fosc	3,9%	17%	1,35
Lighting Conditions	De nit, il·luminació artificial	4,1%	15%	1,20
Lighting Conditions	De nit, il·luminació artificial	11,4%	10%	0,80
Lighting Conditions	De nit, sense llum artificial	6,8%	27%	2,16
Direction of Road	Doble sentit	71,0%	15%	1,22
Direction of Road	Sense especificar	3,1%	4%	0,29
Direction of Road	Un sol sentit	25,9%	9%	0,73

* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

ocean

4.4. Weather patterns of the most severe accidents

The most severe accidents occurs when:
- Surrounding environment: Desmunt, Mixt or Terraple;
- Weather conditions: Pluja debil;
- Wind conditions: vent fort, vent moderat.

Feature	Value	Percent of accidents	Probability of severe	Lift
Fog Presence	No n'hi ha	95,1%	13%	1,01
Fog Presence	Si	4,9%	10%	0,78
Surrounding Environment	A nivell	39,9%	12%	0,95
Surrounding Environment	Desmunt	6,8%	19%	1,53
Surrounding Environment	Mixt	12,4%	20%	1,55
Surrounding Environment	Sense Especificar	37,0%	9%	0,68
Surrounding Environment	Terraplé	3,9%	25%	1,95
Weather Conditions	Bon temps	94,5%	13%	0,99
Weather Conditions	Pluja dèbil	4,2%	15%	1,22
Weather Conditions	Pluja forta	1,2%	13%	1,04
Wind Conditions	Calma, vent molt suau	98,1%	13%	0,99
Wind Conditions	Vent fort	0,4%	22%	1,72
Wind Conditions	Vent moderat	1,5%	15%	1,23

* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

4.5. Involved vehicles patterns of the most severe accidents

- The share of accidents involving heavy vehicles is about 11%, bicycles ~8%, mopeds ~8%, motorcycles ~38%.
- If heavy vehicles is involved in accident then the likelihood of fatalities is greater.
- The opposite is true for bicycles, mopeds and motorcycles.

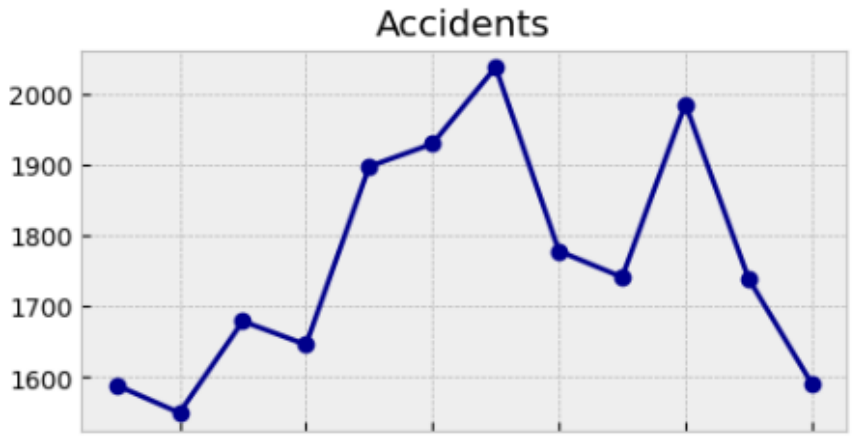
Feature	Value	Percent of accidents	Probability of severe	Lift
Heavy Vehicles Involved	0	88,8%	11%	0,85
Heavy Vehicles Involved	1	10,1%	27%	2,15
Heavy Vehicles Involved	2	1,0%	35%	2,74
Bicycles Involved	0	91,9%	13%	1,04
Bicycles Involved	1	7,5%	7%	0,53
Bicycles Involved	2	0,5%	7%	0,55
Mopeds Involved	0	92,1%	13%	1,06
Mopeds Involved	1	7,7%	4%	0,34
Mopeds Involved	2	0,1%	4%	0,32
Motorcycles Involved	0	61,9%	14%	1,14
Motorcycles Involved	1	36,3%	10%	0,78
Motorcycles Involved	2	1,6%	8%	0,67

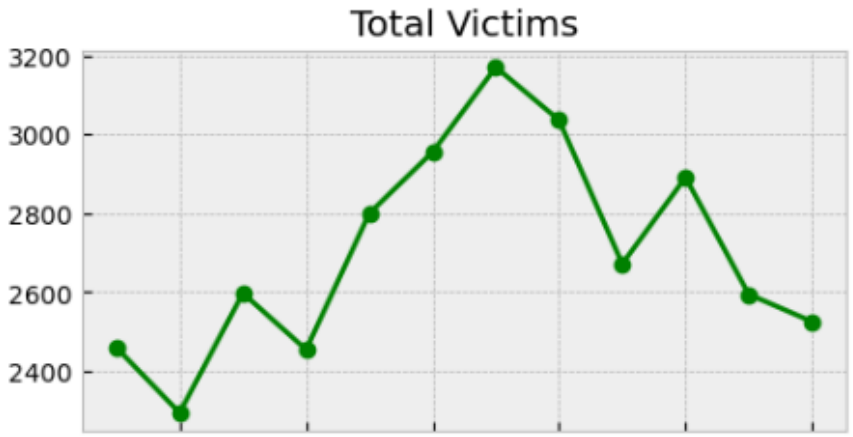
* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

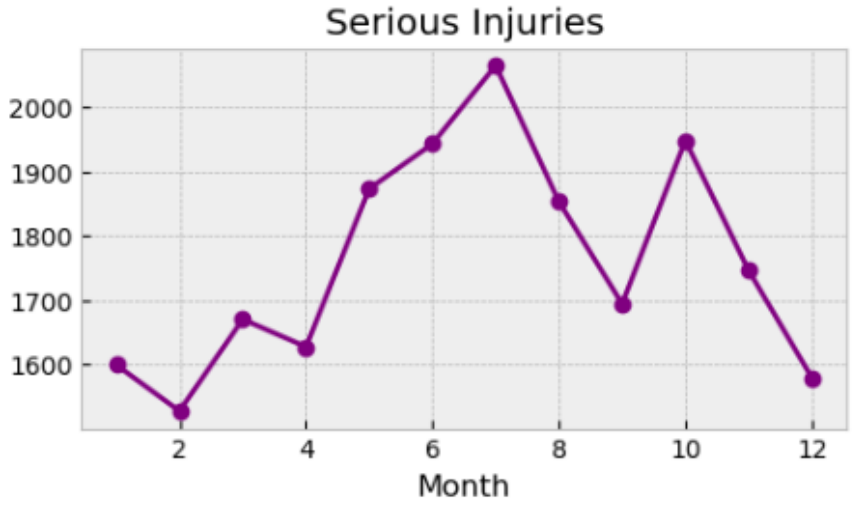
5.1. Accident statistics by month

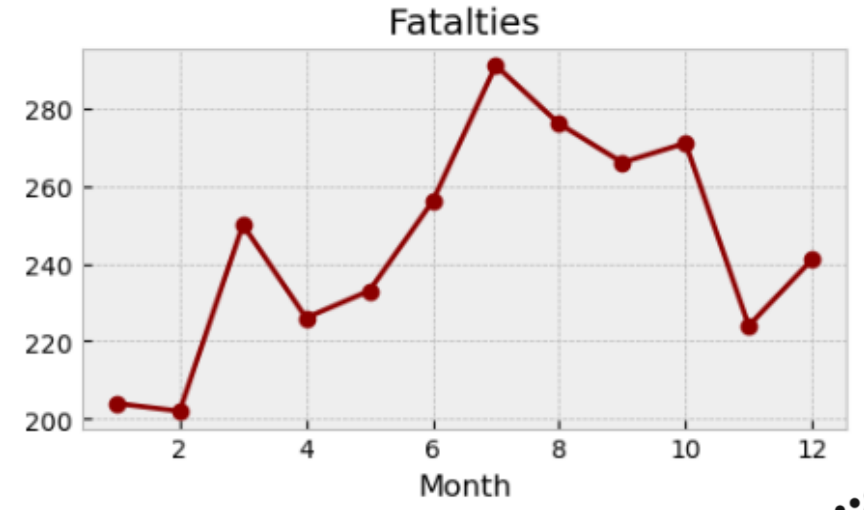
The graphs below show the monthly statistics for the period from 2010 to 2021.

- We see that there are a peak values in July and October.
- In the "extended summer" time (May to October) all statistics have greater values than in the other months.









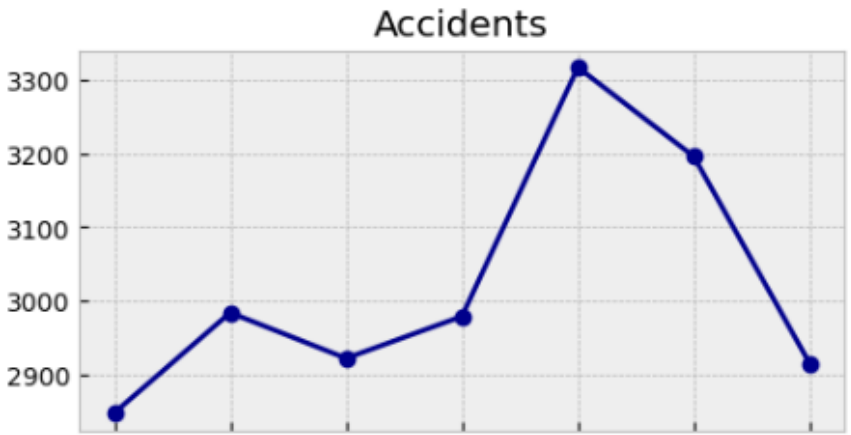
* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

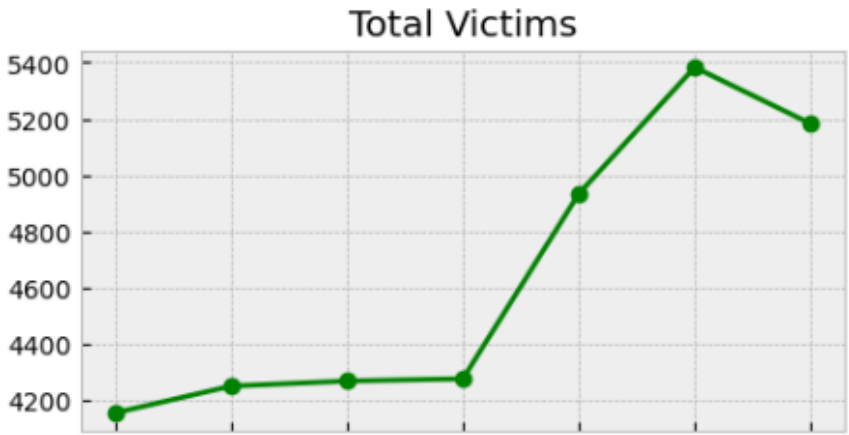
ocean

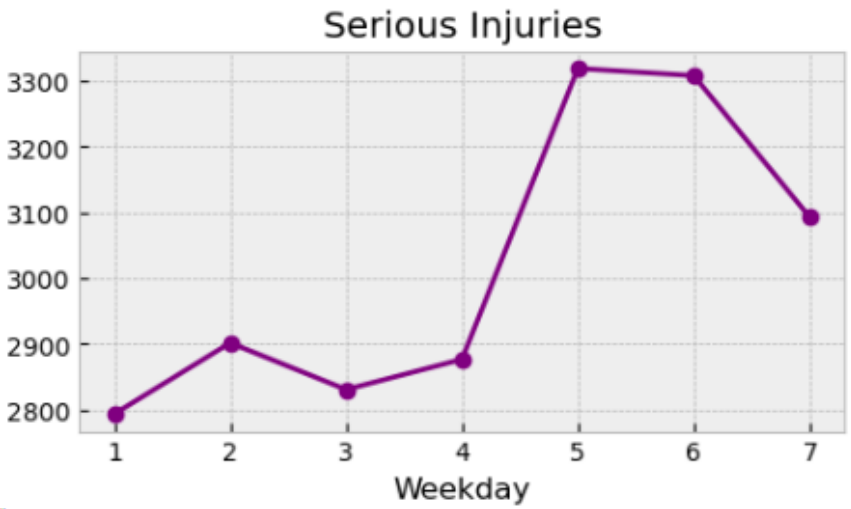
5.2. Accident statistics by the day of week

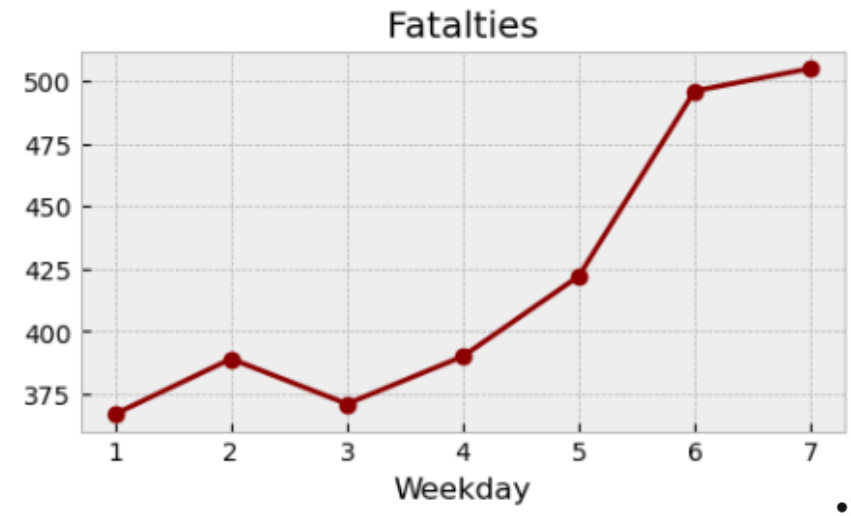
The graphs below show the day of week statistics for the period from 2010 to 2021.

- The most number of accidents are on Friday and Saturday.
- The most number of Total Victims, Serious Injuries and Fatalities are on Friday, Saturday and Sunday, meaning that the average number of victims, serious injuries and deaths per one accident have the highest values on Sunday.
- On weekdays all indicators have relatively low values comparing with weekends.









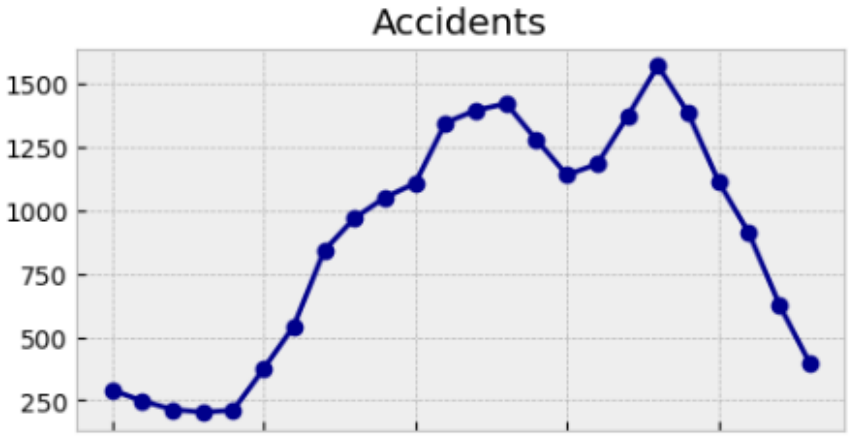
* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

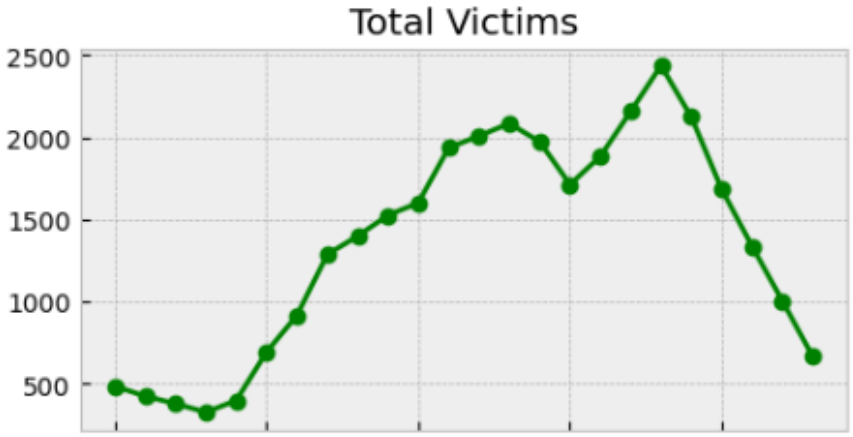
ocean

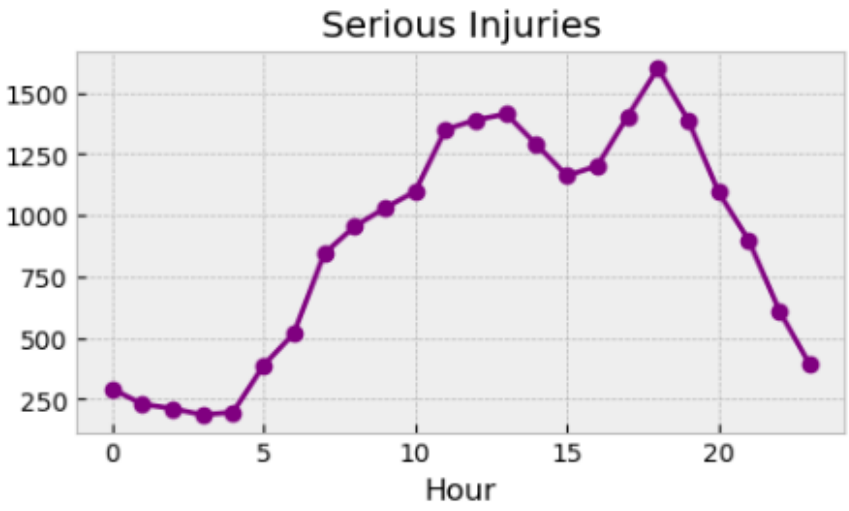
5.3. Accident statistics by the day hours

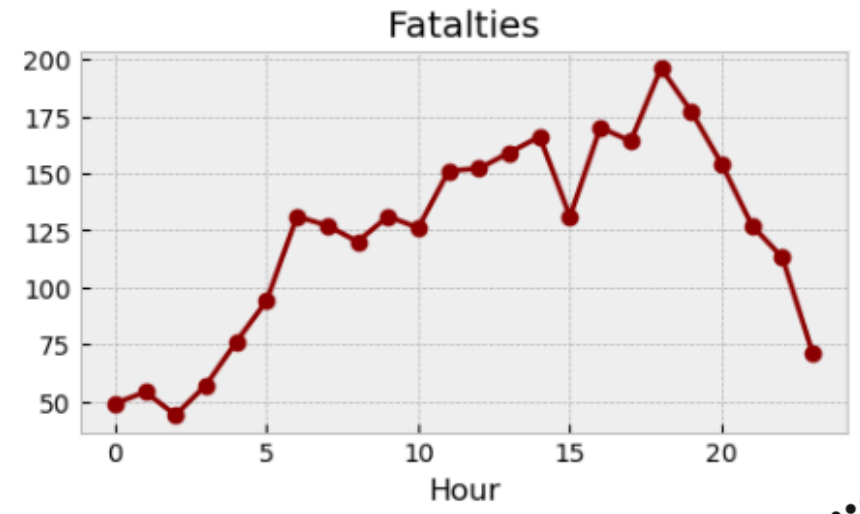
The graphs below show the day hours statistics for the period from 2010 to 2021.

- All indicators have the greatest values in the period from 7:00 to 21:00 with the peak value at 18:00.
- There is an interesting decreasing of indicators from 13:00 to 15:00 and than increasing from 15:00 to 18:00.
- In the night time (from 0:00 to 4:00) all indicators in 3-4 time lower than in the day time from (11:00 to 14:00).









* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

ocean

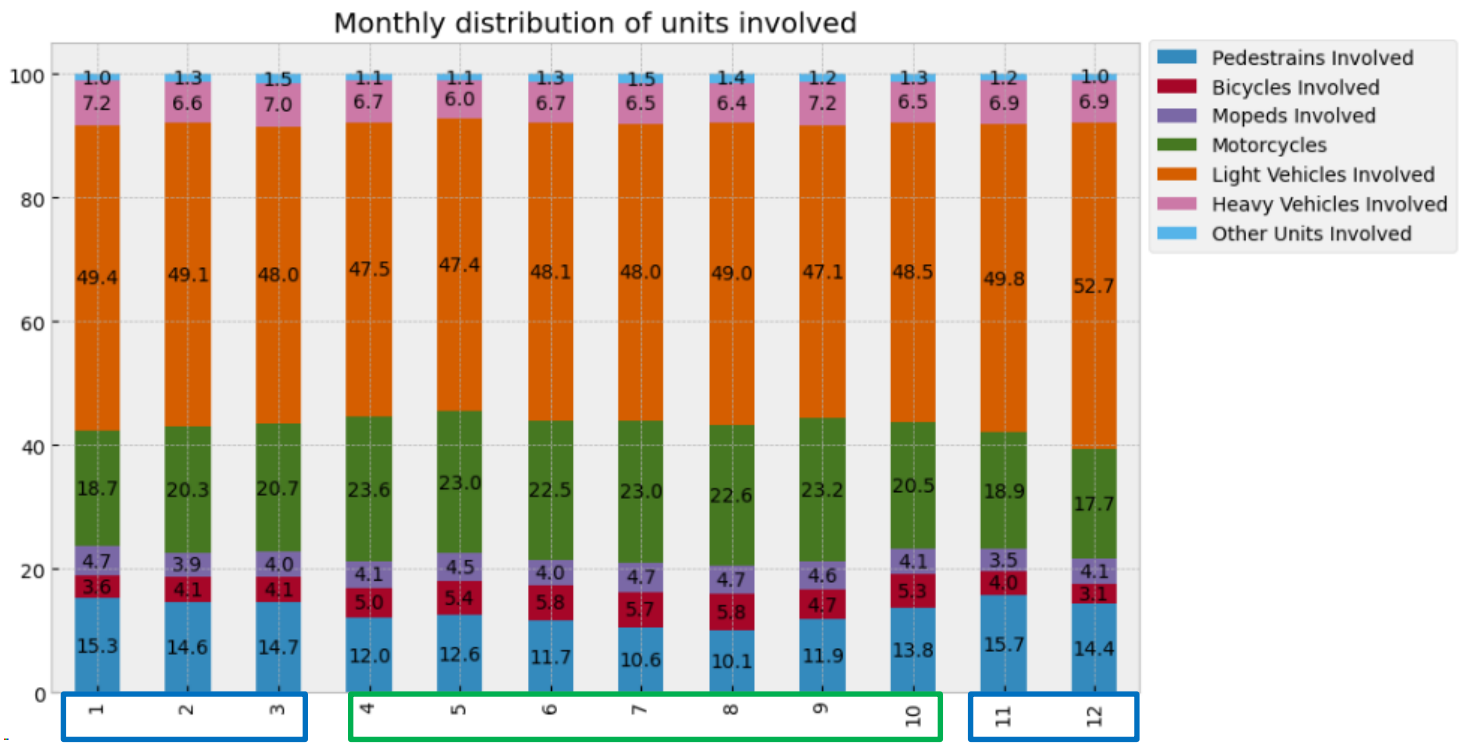
5.4. Monthly clustering by the involved units distribution

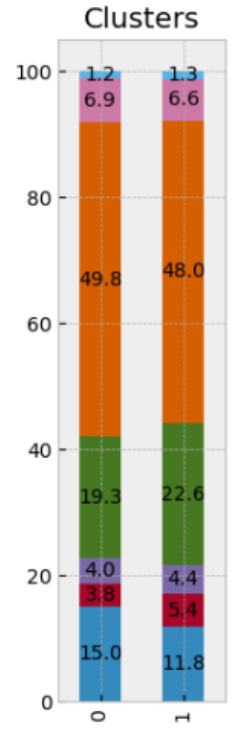
We calculated monthly distributions of different units involved and now want to understand how to split months into groups based on these patterns. By applying K-means clustering algorithm with K = 2 we received two clusters:

- 0: from November to March (winter season);
- 1: from April to October (summer season);

How these clusters differs from each other?

- In the summer season pedestrians share is lower;
- In the summer season bicycles, mopeds and motorcycles share is greater than in the winter season.





* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

ocean

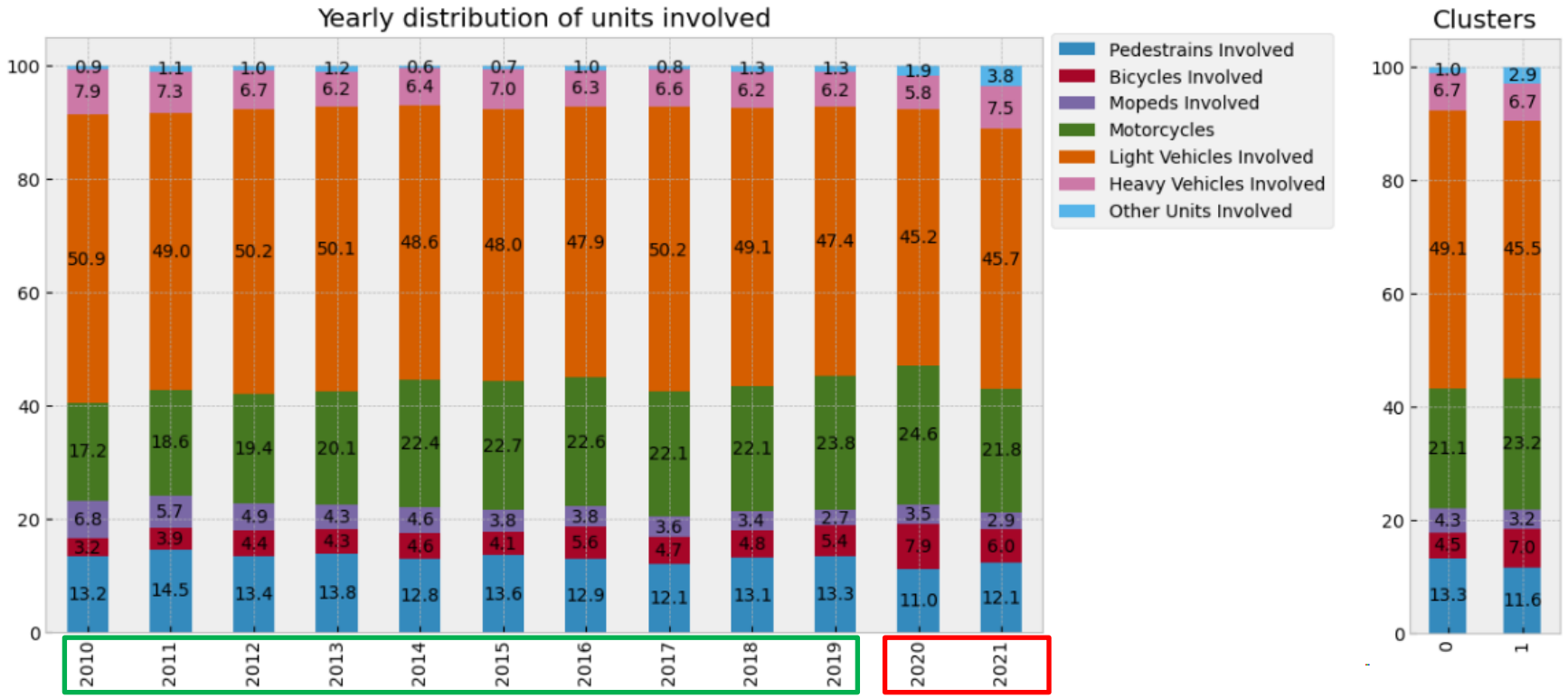
5.5. Yearly clustering by the involved units distribution

We calculated yearly distributions of different units involved and now want to understand how to split years into groups based on these patterns. By applying K-means clustering algorithm with K = 2 we received two clusters:

- 0: from 2010 to 2019 (before COVID-19 time);
- 1: from 2020 to 2021 (COVID-19 pandemic and post-pandemic period).

How these clusters differs from each other?

- In COVID-19 pedestrians and light vehicles shares becomes lower;
- In COVID-19 mopeds share also becomes lower but this is a natural process not related to the pandemic;
- In COVID-19 bicycles and motorcycles shares becomes greater.



* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/1_statistics.ipynb

ocean

6.1. Modeling tasks description

In this section we will develop ARIMA models for predicting yearly number of seriously injured victims in Catalonia.

Dataset preparing:

• As we shown before there was a sharp drop in 2020 due to COVID-19 so we decided to replace the values in 2020 and post-COVID year 2021 by their corresponding neighbors mean values.

• For training and validation we took the data from 2010 to 2022 (the last year is available here: https://www.idescat.cat/indicadors/?id=anuals&n=10753&lang=en&col=1)

Validation:

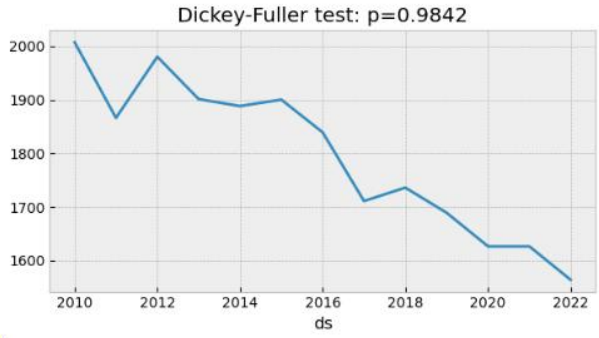
• As the validation set we took the last 2 years: 2022 and 2021.

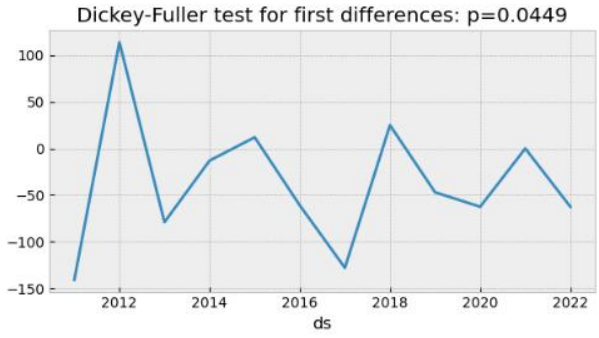
• As a quality metrics we took MAE – mean absolute error.

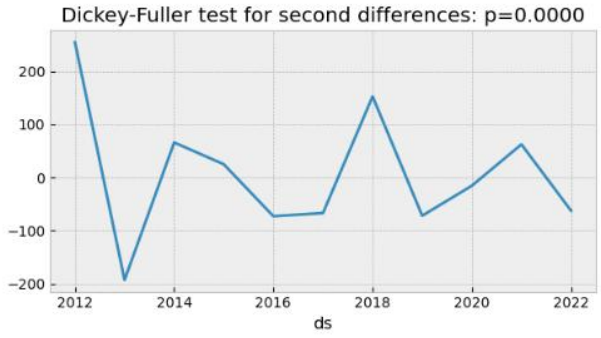
Models searching:

• We searched for the best parameters p, d, q of ARIMA model from the following lists: p in [0,3], d in [0,2], q in [0,3].

• We checked the stationarity of the input series by Dickey-Fuller criterion. The null hypothesis assumes that the process is non-stationary, the alternative hypothesis says the opposite. As we can see on the picture below we can significantly reject the null hypothesis only for the second differences of initial series meaning that our best models should have d = 2. We check that later.







* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/2_modeling.ipynb

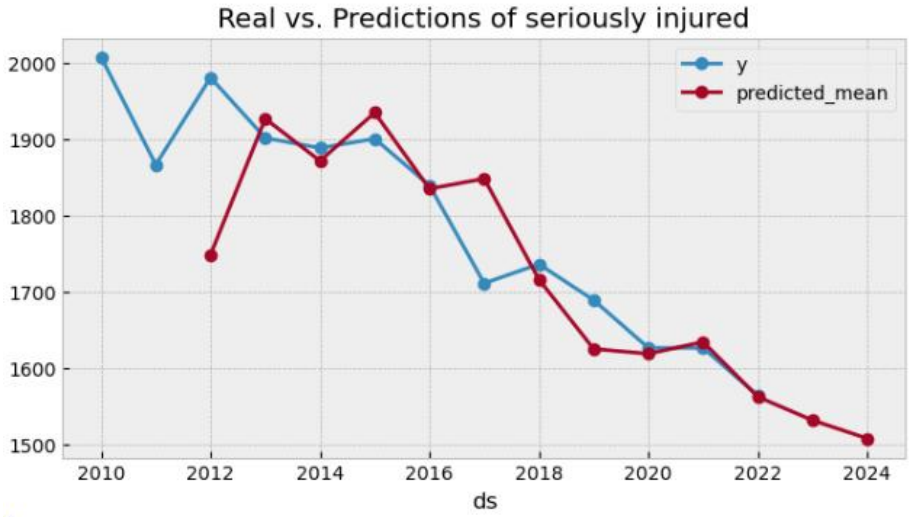
ocean

6.2. ARIMA modeling results

- As we see in the table below that our best model ARIMA(2,2,0) have **MAE~1.3** which is very good result**.**
- Below we shown the best model predictions on the validation set and upcoming 2023 and 2024 years.
- In the table below we shown all coefficients of best model that as we see all are significant.

Best model searching:

	p	d	q	mae
0	2	2	0	4.79
1	3	2	1	8.18
2	3	1	1	9.92
3	1	2	1	10.16
4	2	1	1	11.50
5	2	2	1	11.72
6	0	2	2	14.27
7	0	2	3	15.52
8	3	2	0	18.31
9	3	2	3	18.75



Best model description:

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	13
Model:	SARIMAX(2, 2, 0)	**Log Likelihood**	-61.559
Date:	Tue, 30 Jan 2024	**AIC**	129.118
Time:	23:08:46	**BIC**	130.312
Sample:	0	**HQIC**	128.366
	- 13		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.2020	0.206	-5.825	0.000	-1.606	-0.798
ar.L2	-0.7174	0.326	-2.198	0.028	-1.357	-0.078
sigma2	3713.6322	1795.127	2.069	0.039	195.249	7232.016

Ljung-Box (L1) (Q):	0.19	**Jarque-Bera (JB):**	1.09
Prob(Q):	0.66	**Prob(JB):**	0.58
Heteroskedasticity (H):	0.28	**Skew:**	-0.50
Prob(H) (two-sided):	0.24	**Kurtosis:**	4.18

* More details you can find in the script: https://github.com/abessalov/Ocean_Traffic/blob/master/2_modeling.ipynb

ocean