ocean

PROJECT REPORT

Road to Safety: Traffic Accident Analysis

Year : 2024

Analyst : Hugo

ABSTRACT

In this report, we present an in-depth analysis of a critical dataset from the Road to Safety: Traffic Accident Analysis Data Challenge, a key component of the 2024 Data Challenge Championship. This dataset encompasses a comprehensive record of traffic accidents in Catalonia from 2010 to 2021, including vital data on collisions, fatalities, and serious injuries. The analysis leverages advanced data visualization and statistical analysis techniques to extract meaningful insights and patterns, aiming to enhance understanding of road safety issues. Through meticulous examination, the report reveals significant trends and contributing factors to road accidents, providing a foundation for actionable strategies to improve traffic safety. This study not only contributes to the field of data science but also offers valuable perspectives for predictive modeling and policy formulation in road safety, employing cutting-edge deep learning methodologies.

KEY WORDS

Traffic Accident Analysis, Data Science, Deep Learning, Road Safety, Statistical Analysis.

Table of Contents

Introduction **4**

1. Data Preparation **5**

a. Missing values 5

b. Data Encoding 6

c. Outliers and Correlation Analysis 7

2. Data Visualizations **8**

a. What are the overall trends in traffic accidents, fatalities, and serious injuries in Catalonia from 2010-2021? 9

b. What common characteristics (time of day, type of road, etc.) are observed in the most severe accidents? 10

c. Which municipalities or counties in Catalonia have the highest incidence of traffic accidents? How does this correlate with population density or road network characteristics? 11

d. How have traffic accident patterns (frequency, severity) changed yearly from 2010 to 2021? 13

e. On what days of the week and times of day do most accidents occur? Are there notable differences between weekdays and weekends? 15

f. How do different weather conditions affect the likelihood of accidents? Is there a correlation between visibility, road conditions, and accident severity? 17

g. What impact do road features (such as speed limits and road types) and traffic density have on the occurrence of accidents? 18

h. Does the involvement of specific types of vehicles (like heavy trucks and motorcycles) correlate with more severe accidents? 20

i. Are there specific periods (months, years) where accident patterns cluster significantly? What might be the causes for these clusters? 22

3. Model and Prediction **24**

a. Time Series Modeling 24

b. Deep Learning Approach 26

c. Evaluation 26

Conclusion **28**

Appendix **29**

Image related to “What common characteristics (time of day, type of road, etc.) are observed in the most severe accidents?” 29

Image related to “How do different weather conditions affect the likelihood of accidents? Is there a correlation between visibility, road conditions, and accident severity?” 31

Code 33

Glossary

Algorithms: A set of rules or instructions given to an AI, computer program, or model to help it learn and make decisions.
Data Preparation: The process of cleaning and organizing raw data into a suitable format for analysis or modeling.
Data Analysis: The process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.
Categorical Data: Data that can be divided into specific groups or categories but does not have a natural order.
Model: In machine learning, a model is the output of a training process and is used for making predictions.
Deep Learning: A subset of machine learning involving neural networks with many layers, particularly effective for complex tasks like image and speech recognition.
Mapping: The process of transforming data elements from one structure to another, often used in data visualization.
DateTime: In programming, a type of data that includes date and time information.
Timestamp: A sequence of characters or encoded information identifying when a certain event occurred, usually giving date and time of day.
Correlation Matrix: A table showing correlation coefficients between variables, used to examine relationships within data.
Normalization: The process of scaling individual data points to have a more uniform scale without distorting differences in the ranges of values.
Long Short-Term Memory (LSTM): A type of recurrent neural network used in deep learning, capable of learning order dependence in sequence prediction problems.
Neural Networks: Computational models inspired by the human brain, consisting of interconnected nodes or neurons that process information using a connectionist approach.
ARIMA (AutoRegressive Integrated Moving Average): A popular statistical analysis model used for time-series forecasting.
AutoRegression: A model that uses the dependent relationship between an observation and a number of lagged observations.
Regression: In statistics, it's a predictive modeling technique for estimating relationships among variables.

Introduction

This report delves into a comprehensive analysis of traffic accident data in Catalonia from 2010 to 2021. Our goal is to uncover patterns and insights that could contribute to enhancing road safety. We aim to present these findings in a manner that is easily understandable, steering clear of technical jargon and focusing on clarity and engagement.

The analysis focuses on various aspects of traffic incidents, including collisions, fatalities, and serious injuries, with an intent to identify underlying factors that contribute to these occurrences. Through this report, we endeavor to communicate key insights that are not only significant for road safety authorities and policymakers in Catalonia but also offer valuable lessons for global road safety measures.

We leverage clear and intuitive data visualizations to supplement our findings, ensuring that our charts and graphs are straightforward and self-explanatory. This approach is intended to make the data more accessible and the insights more apparent to our readers.

Our in-depth analysis is structured to provide clear and compelling explanations of our findings. Wherever possible, we use analogies and simple comparisons to make complex data more relatable and understandable.

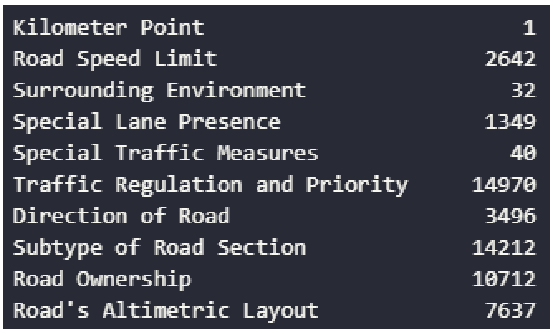
In conclusion, we aim to summarize the critical takeaways from our analysis, discussing their implications and suggesting actionable steps based on our discoveries. We hope that this report serves as a valuable tool in the ongoing efforts to improve road safety and prevent traffic-related incidents.

1. Data Preparation

Data preparation is a crucial stage in any data analysis process, often considered the foundation upon which reliable and meaningful insights are built. In our analysis of the Catalonia traffic accident dataset, extensive data preparation was undertaken to ensure the accuracy and integrity of the findings. This section outlines the key steps in our data preparation process, emphasizing their importance in enhancing the quality of the final analysis.

a. Missing values

The first step involved addressing missing data, a common issue in real-world datasets.



```
Kilometer Point                        1
Road Speed Limit                    2642
Surrounding Environment               32
Special Lane Presence               1349
Special Traffic Measures              40
Traffic Regulation and Priority    14970
Direction of Road                   3496
Subtype of Road Section            14212
Road Ownership                     10712
Road's Altimetric Layout            7637
```

1. Number of missing values

Addressing missing data is a fundamental step in data preparation, pivotal for ensuring the accuracy and reliability of any subsequent analysis. But first, let's understand what constitutes a missing value and its implications. In our dataset, consisting of 21,161 rows and 58 columns, missing values are quite common. These can occur due to various reasons - sometimes data points are inadvertently omitted during entry, or certain information might just not be available.

Missing values can lead to biased or inaccurate analyses if not handled properly. They represent gaps in our data that can distort statistical calculations and compromise the validity of the analysis. The challenge is to address these missing values without compromising the integrity of the dataset.

To manage these missing values, we choose three strategies. When a column has very few missing values, one approach is to delete the rows containing these missing values. This method, while not always ideal, can sometimes be a practical solution to prevent the distortion of statistical analyses. It’s a trade-off between retaining the maximum amount of data and ensuring the quality of the dataset. In cases where missing values can be logically deduced, we utilize external knowledge to fill in the gaps. A prime example in our dataset is the missing speed limit data. With 2,642 missing values, we turned to Spanish traffic laws, using the known road types to infer the corresponding speed limits. This method allows for a logical and informed approach to filling missing values, thereby maintaining the dataset’s utility without introducing significant biases. Sometimes, a missing value is an indication of the absence of data itself, and in such cases, we can label these as 'Not indicated.' This approach acknowledges the lack of data while ensuring that its absence doesn’t skew the analysis.

The decision on how to handle missing values is context-specific and hinges on understanding the dataset and the nature of the missing data. Each approach has its merits and drawbacks, and the choice largely depends on the extent and impact of the missing data on the overall dataset.

In summary, our approach to managing missing data was multi-faceted, considering the nature of each missing value and the potential impact of different handling strategies on the dataset as a whole. We aimed to strike a balance between preserving data integrity and maintaining the robustness of our analysis. This meticulous handling of missing data not only ensured the reliability of our dataset but also laid a strong foundation for the accuracy of our insights and the effectiveness of our predictive models.

By thoughtfully addressing missing values through deletion, informed imputation, or clear indication of data absence, we enhanced the quality of our dataset, thereby ensuring that our analysis was grounded in the most complete and accurate information available. This process exemplifies the importance of careful data preparation in any data-driven endeavor, particularly in a domain as impactful as traffic safety analysis.

b. Data Encoding

In the field of data analysis and deep learning, correctly managing and encoding different data types is crucial for effective model performance and accurate insights. This section of the report is dedicated to explaining why and how we handled data encoding in our analysis of the traffic accident dataset.

Many deep learning algorithms can only interpret numerical values. Thus, categorical data, which often come in textual form (like names of streets or types of vehicles), must be converted into a numerical format to be processed effectively by these algorithms.

Proper encoding ensures that the algorithms correctly understand the nature and hierarchy of the data. For instance, ordinal data have a specific order that must be preserved, which requires a different encoding approach compared to nominal data where no such order exists.

Correct data encoding can significantly improve the efficiency and performance of deep learning models. It can reduce memory usage and speed up the processing of data.

The first step was to identify different data types in our dataset. We focused on distinguishing between numerical, categorical (nominal and ordinal), and datetime data. This identification was crucial for deciding the encoding techniques to be applied. To define a comprehensible and easy-to-use encoding, we used mapping. This allows us to assign numbers to categorical values easily.

DateTime data were processed to extract meaningful features that could significantly impact the analysis. For instance, extracting the day of the week, month, or time of the day from a timestamp can reveal trends and patterns related to traffic accidents.

In conclusion, the process of data encoding was a critical component in preparing our traffic accident dataset for analysis. By transforming categorical and datetime data into appropriate numerical formats, we were able to facilitate more accurate and efficient data processing, which is key for any data-driven analysis, especially in deep learning contexts. This careful and methodical approach to data encoding not only enhanced the quality of our dataset but also played a pivotal role in the success of our subsequent analytical processes.

Through this encoding process, we ensured that the nuances and specific characteristics of our data were accurately captured and represented in a format that is suitable for advanced analytical techniques. This step was instrumental in uncovering the valuable insights that were essential for our goal of enhancing road safety and understanding traffic accident patterns.

c. Outliers and Correlation Analysis

In our data preparation process for the traffic accident dataset, special attention was given to managing outliers and understanding correlations between different variables. This section details our approach to these two crucial aspects of data analysis

Outliers are data points that significantly differ from other observations and can skew our analysis. In our dataset, a careful examination revealed that there were no significant outliers that warranted removal or adjustment. This absence of outliers is crucial as it indicates the data's consistency and reliability.

The identification of outliers was carried out using statistical methods with mean, max and min values with the proportion. The consistency in data patterns suggested that most data points were within a reasonable range of values, typical for traffic accident data.

The lack of significant outliers implies that the dataset provides a realistic representation of traffic accidents without extreme anomalies. This consistency ensures that the models and analyses developed will be more representative of typical conditions, leading to more reliable and applicable results.

Correlation analysis helps in understanding the relationship between different variables. In our dataset, a correlation matrix was utilized to identify relationships between various factors. A notable finding was the strong correlation between 'Total Victims' and 'Light Injuries.' This correlation suggests that a high percentage of victims in traffic accidents suffer from light injuries.

Based on our analysis, we concluded that the 'Total Victims' column could be removed. This decision is underpinned by the observation that the sum of different types of injuries (light, serious, etc.) should logically equate to the total number of victims. Removing the 'Total Victims' column helps to simplify the dataset and focus on more specific and descriptive variables, like the types of injuries, which provide more granular insights into the nature of the accidents.

Removing a highly correlated variable reduces redundancy in the dataset. It helps in preventing multicollinearity issues in our predictive models, where independent variables are highly correlated. This step ensures that our models are more robust and the interpretations of the model’s outputs are clearer and more reliable.

In conclusion, the careful management of outliers and the strategic decision to remove the 'Total Victims' column based on correlation analysis were pivotal in enhancing the dataset's quality. These steps ensured that our analysis was focused, relevant, and grounded in a clear understanding of the relationships between different variables, thereby paving the way for more accurate and insightful findings.

2. Data Visualizations

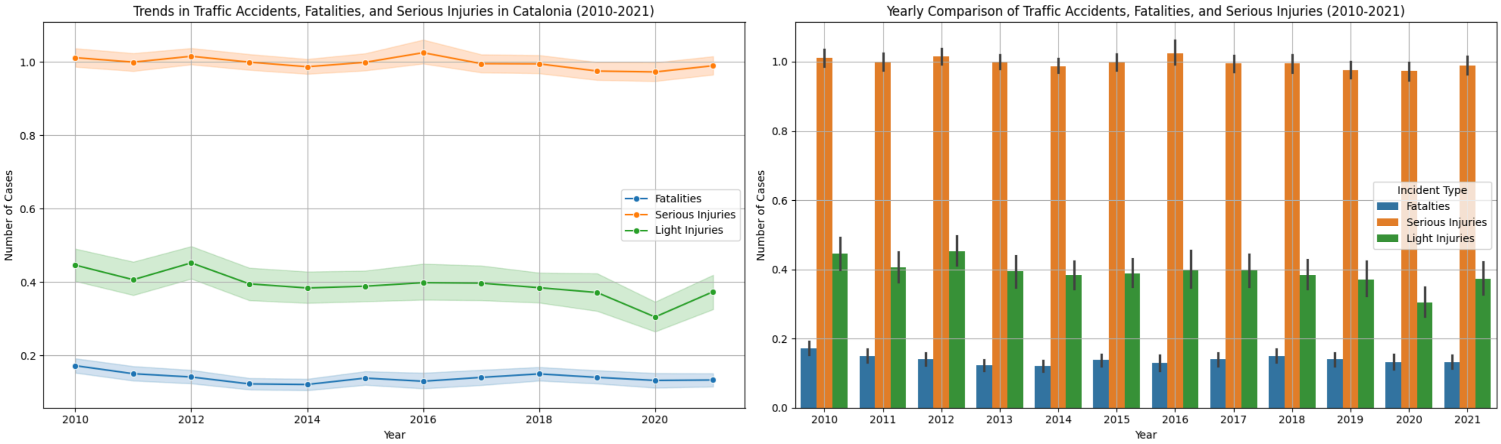
The data visualization section of our report plays a crucial role in the elucidation of complex data and the communication of our findings. In an age where data is abundant, the ability to distill vast amounts of information into clear, concise, and engaging visual narratives is indispensable. Data visualizations do more than just present numbers; they bring data to life, uncovering patterns and insights that might remain hidden in spreadsheets or textual analysis.

Visualizations leverage the human visual system's ability to identify trends, differences, and outliers quickly. By transforming data into graphical representations, we can see the shape of data, compare various elements, and track changes over time at a glance. This immediacy of comprehension is invaluable, particularly when dealing with multidimensional data that spans over a decade, as is the case with our traffic accident dataset for Catalonia.

In this section, we employ a variety of visualization techniques, each tailored to the specific type of data and analysis. Bar charts provide a clear comparison of categorical data, such as the number of accidents by type of road or time of day, allowing us to instantly grasp which categories are most prevalent. Line graphs depict trends over time, showing how variables have increased, decreased, or fluctuated throughout the years. Heatmaps offer a powerful means of understanding the relationship between two variables, revealing density and concentration that would be difficult to discern from raw numbers alone.

The importance of data visualization is not merely in aesthetic appeal but in its capacity to inform, persuade, and drive decision-making. A well-crafted visualization can highlight the most important aspects of the data, guiding the viewer to key insights and supporting more informed policy and decisions. As we navigate through the visual stories in this section, we aim to provide a clear window into the dynamics of traffic accidents in Catalonia, drawing attention to critical areas and times where interventions might be most needed to improve road safety and save lives.

a. What are the overall trends in traffic accidents, fatalities, and serious injuries in Catalonia from 2010-2021?



2. Illustration of the trends in traffic accidents, fatalities, and serious injuries in Catalonia over the period from 2010 to 2021.

The line graph illustrates trends over the years. Minor injuries appear to be a frequent consequence of traffic incidents, their frequency remaining relatively stable, with a slight increase towards the end of the period. Fatalities, although the least frequent, show no significant increase or decrease over time, maintaining a relatively stable trend. Serious injuries, which are clearly more prevalent than others, show a slight upward trend, particularly in the last few years of the period.

The bar graph provides a year-by-year comparison of the incidents. Each year is depicted with three bars representing fatalities, serious injuries, and light injuries, respectively. The bars corresponding to serious injuries are consistently the tallest, reiterating that they occur with the greatest frequency. The bars for light injuries and fatalities are shorter, suggesting these are less common outcomes. The relative proportions between the three types of incidents remain roughly consistent throughout the years, though there is some variation.

These visualizations complement one another, offering insights into both the individual and relative changes in traffic incidents over the twelve-year span. The line graph allows for the observation of long-term trends and trajectories, while the bar graph affords a more immediate comparison between the years, making it simple to identify any peculiar year-to-year changes. Together, they provide a clear and informative picture of the traffic accident landscape in Catalonia, underlining the overarching trends in road safety and incident occurrence.

b. What common characteristics (time of day, type of road, etc.) are observed in the most severe accidents?

Project report - Ocean Protocol

The visual analysis of the data visualizations (appendix) reveals a multifaceted portrait of severe accidents. When we look at the times of day, there's a noticeable concentration of both fatal and serious accidents in the afternoon and morning hours, with a marked decrease during the night. This trend points to the busier traffic conditions of daylight hours as a significant factor in the risk of severe accidents.

Delving into the types of roads where these accidents occur, we find that conventional roads and urban streets are the most common settings for severe incidents. These environments likely present a complex array of challenges for drivers, from navigating intersections to sharing the road with pedestrians and cyclists, which could account for the higher incidence of accidents.

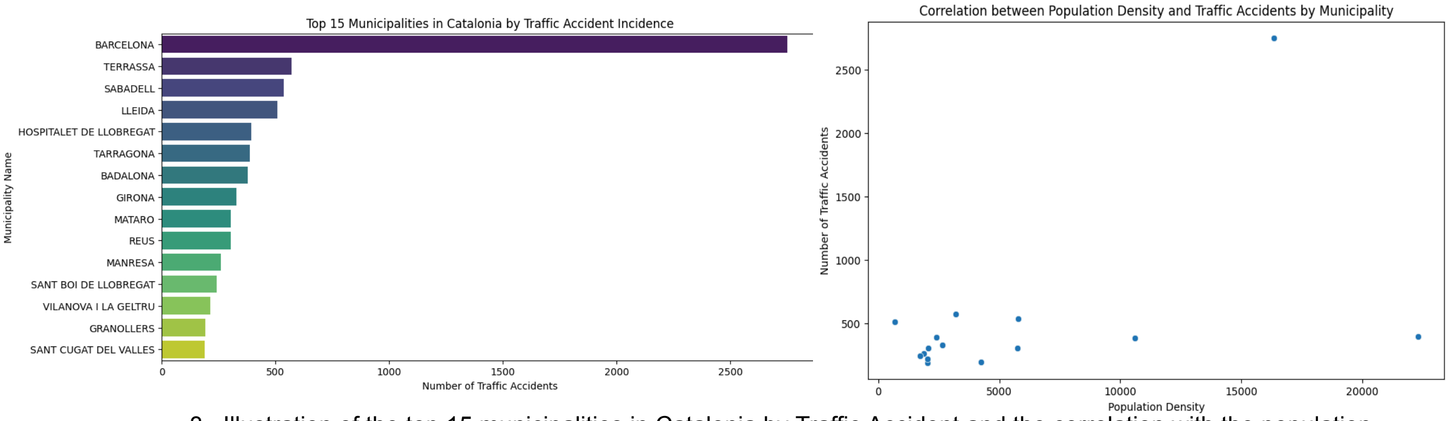
Speed limits also emerge as a critical factor. The data shows a pronounced peak in fatal accidents at the 100 km/h limit, suggesting that accidents occurring at higher speeds are more likely to result in fatalities. This is a clear signal of the dangers of high-speed driving and underscores the importance of enforcing speed limits for road safety.

Looking at the broader picture, the area in which accidents occur also shapes the severity. Roads see a higher number of fatal accidents than urban areas, while urban areas record more serious accidents. This could reflect the variance in driving conditions, such as higher speeds on open roads versus the more congested, but slower-paced, urban settings.

Across months and days, the data does not indicate any significant variation, suggesting that the risk of severe accidents is relatively uniform throughout the year and across the calendar. The consistency across these timeframes indicates that while certain times and settings are more prone to severe accidents, the potential for serious incidents is ever-present, necessitating constant vigilance and safety measures.

These insights, drawn from a careful examination of the visual data, highlight the complexity of factors that contribute to severe traffic accidents. Understanding these patterns is essential for developing targeted, effective strategies for improving road safety and reducing the occurrence of these tragic events.

c. Which municipalities or counties in Catalonia have the highest incidence of traffic accidents? How does this correlate with population density or road network characteristics?



3. Illustration of the top 15 municipalities in Catalonia by Traffic Accident and the correlation with the population density

Using the visualizations above, we can see that Barcelona, Terrassa, and Sabadell are the municipalities in Catalonia with the highest incidence of traffic accidents. Barcelona stands out significantly, with more than double the number of accidents compared to any other municipality. When correlating this with population density, there is an evident trend: higher population density tends to be associated with a higher number of traffic accidents. The bar chart showing the "Top 15 Municipalities in Catalonia by Traffic Accident Incidence" suggests that more populous areas like Barcelona, which have a high density of inhabitants and likely a more extensive road network, experience a greater number of accidents.

FRANCE
ANDORRE
ARAGÓN
VALENCIA
MER MÉDITERRANÉE
Costa Daurada
Costa Brava
NORD
Vall d'Aran
Vielha
Perpignan
Puigcerdà
Llívia
La Seu d'Urgell
Tremp
Lleida
Balaguer
Cervera
Manresa
Vic
Ripoll
Olot
Figueres
Gerona
Cadaqués
Calella de Palafrugell
Sant Feliu de Guíxols
Tossa de Mar
Lloret de Mar
Blanes
Mataró
Granollers
Terrassa
Sabadell
Badalona
BARCELONA
Sitges
Vilafranca del Penedès
Tarragona
Port Aventura
Reus
Valls
Tortosa
Amposta
Delta de l'Ebre
Vinaròs
Peníscola
Fraga
Monzón
Barbastro
Ainsa
0
20
40 km
© Cartographie HACHETTE Tourisme

4. Road map Catalunya

Moreover, an examination of the road network in Catalonia reveals that Barcelona serves as a nexus for major transportation routes, including the primary freeways. Considering that Barcelona's population swells significantly during the summer and other holiday periods due to tourism, this influx can readily explain the spike in accident numbers. The increase is further compounded as many tourists opt for bicycles, taxis, or rental cars for mobility within the city, often influenced by the relatively high cost of metro travel for those without a long-term pass.

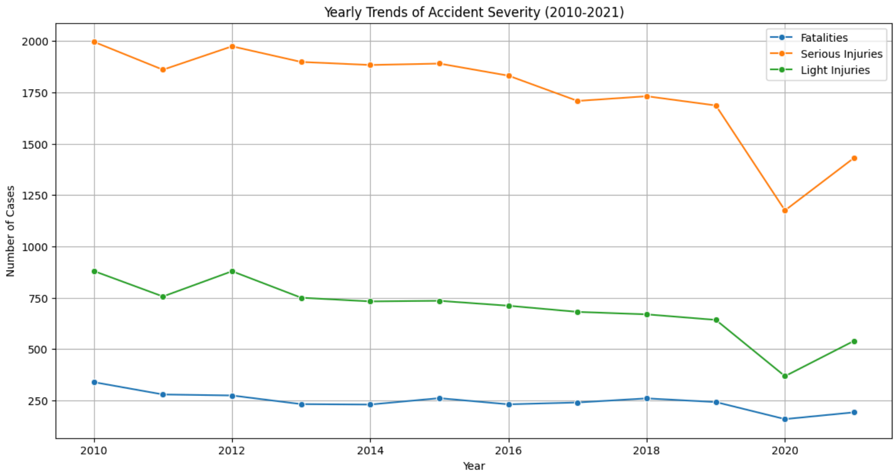
The scatter plot "Correlation between Population Density and Traffic Accidents by Municipality" further illustrates this relationship. Municipalities with higher population densities, such as Hospitalet de Llobregat and Barcelona, also exhibit a higher number of traffic accidents. This correlation suggests that as the number of people living in a confined area increases, so does the likelihood of traffic accidents, potentially due to factors such as increased vehicular traffic, the complexity of the road network, and the frequency of interactions among road users.

However, the correlation is not perfect. Some municipalities with high population densities do not have as many accidents as might be expected, and vice versa. This indicates that other factors, such as the quality of road infrastructure, traffic management systems, and local driving behaviors, might also play significant roles in influencing the incidence of traffic accidents.

To provide a more accurate analysis, additional data would be needed, such as the exact road network characteristics, including road types, condition, and traffic volumes, as well as other potentially influential factors like the presence of public transportation options, the effectiveness of traffic law enforcement, and the general urban design of each municipality.

Overall, while there is a clear pattern that larger, denser municipalities tend to have a higher incidence of traffic accidents, the full picture of what influences traffic safety is complex and multifaceted. Comprehensive data analysis is crucial for urban planners and policymakers to understand the underlying causes and implement effective measures to improve road safety.

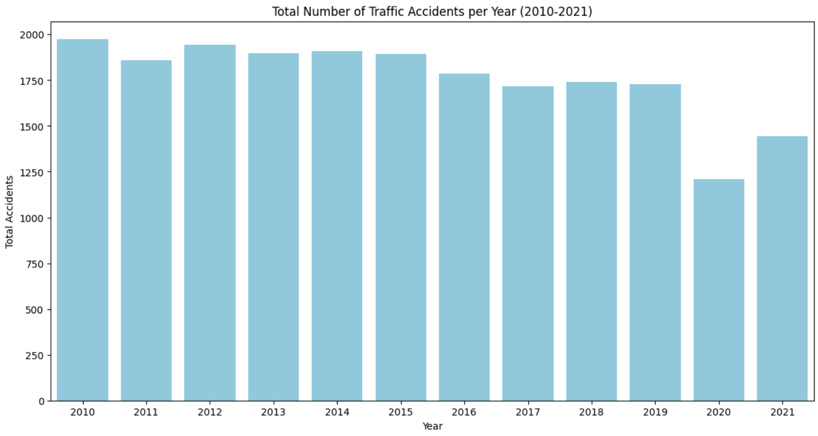
d. How have traffic accident patterns (frequency, severity) changed yearly from 2010 to 2021?



5. Line chart of Yearly Trends of Accident Severity

From the "Yearly Trends of Accident Severity" line chart, we observe the following patterns:

Fatalities: The number of fatalities appears relatively stable from 2010 until around 2018, with slight fluctuations. However, there is a noticeable drop in 2019, followed by a sharp decrease in 2020. This decline could be attributed to various factors, including improved road safety measures or reduced travel during the COVID-19 pandemic. In 2021, there is a rebound, which could indicate a return to pre-pandemic conditions.



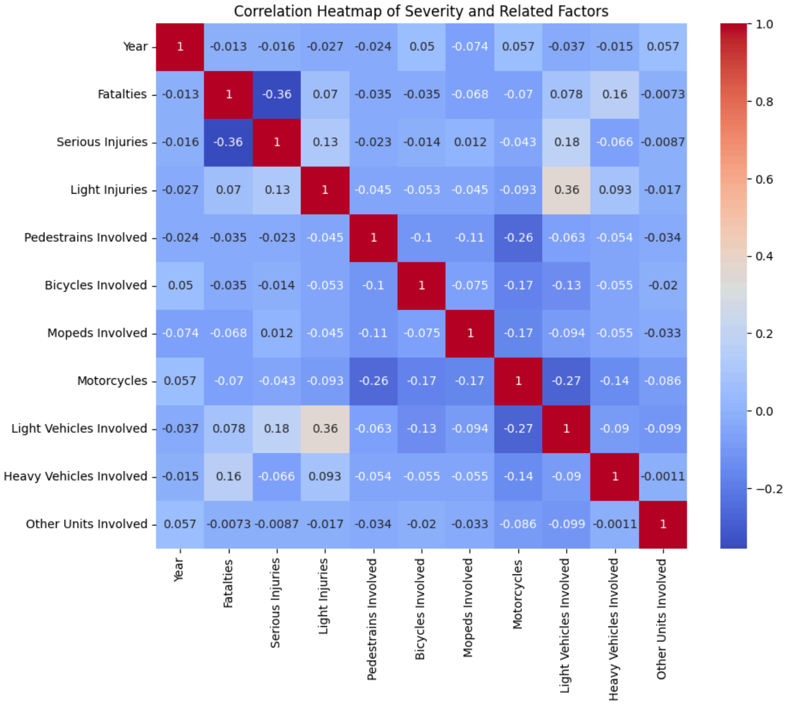
6. Bar chart of Total Number of Traffic Accidents per Year

Serious Injuries: The trend for serious injuries displays some variance over the years but remains within a specific range without any drastic changes. There is a gentle decrease over the years, which could suggest gradual improvements in vehicle safety technology and emergency response.

Light Injuries: The number of light injuries shows a decrease over the years, with a significant drop in 2020, much like the pattern seen with fatalities. This again could be due to reduced travel during the pandemic or better safety practices.

The "Total Number of Traffic Accidents per Year" bar chart shows a consistent decrease in the total number of accidents from 2010 to 2021, with the most significant drop occurring in 2020. This downward trend suggests that overall, the frequency of traffic accidents has been decreasing over the years.

The "Correlation Heatmap of Severity and Related Factors" provides insight into how various factors are associated with the severity of accidents. From the heatmap, we can infer that there is no strong correlation between the year and the severity of accidents, indicating that the change in patterns over the years is not heavily influenced by the factors listed in the heatmap. Instead, the heatmap shows stronger correlations within specific types of involvement, such as between 'Light Vehicles Involved' and 'Light Injuries', which is to be expected as more light vehicles involved in accidents could lead to more light injuries.

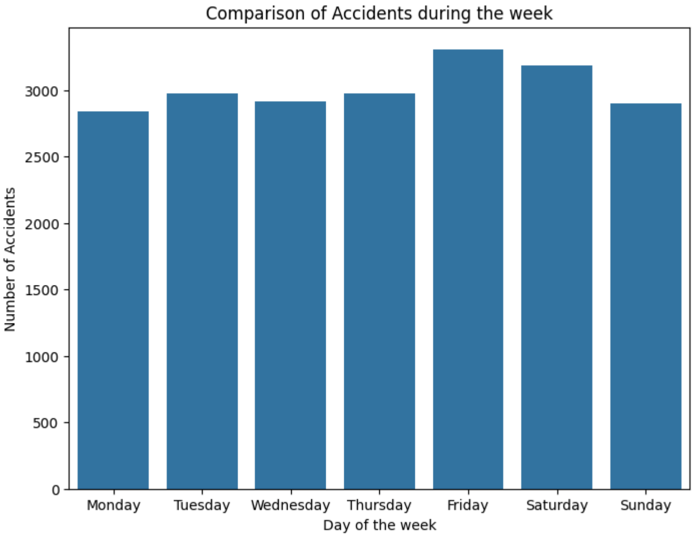


	Year	Fatalties	Serious Injuries	Light Injuries	Pedestrains Involved	Bicycles Involved	Mopeds Involved	Motorcycles	Light Vehicles Involved	Heavy Vehicles Involved	Other Units Involved
Year	1	-0.013	-0.016	-0.027	-0.024	0.05	-0.074	0.057	-0.037	-0.015	0.057
Fatalties	-0.013	1	-0.36	0.07	-0.035	-0.035	-0.068	-0.07	0.078	0.16	-0.0073
Serious Injuries	-0.016	-0.36	1	0.13	-0.023	-0.014	0.012	-0.043	0.18	-0.066	-0.0087
Light Injuries	-0.027	0.07	0.13	1	-0.045	-0.053	-0.045	-0.093	0.36	0.093	-0.017
Pedestrains Involved	-0.024	-0.035	-0.023	-0.045	1	-0.1	-0.11	-0.26	-0.063	-0.054	-0.034
Bicycles Involved	0.05	-0.035	-0.014	-0.053	-0.1	1	-0.075	-0.17	-0.13	-0.055	-0.02
Mopeds Involved	-0.074	-0.068	0.012	-0.045	-0.11	-0.075	1	-0.17	-0.094	-0.055	-0.033
Motorcycles	0.057	-0.07	-0.043	-0.093	-0.26	-0.17	-0.17	1	-0.27	-0.14	-0.086
Light Vehicles Involved	-0.037	0.078	0.18	0.36	-0.063	-0.13	-0.094	-0.27	1	-0.09	-0.099
Heavy Vehicles Involved	-0.015	0.16	-0.066	0.093	-0.054	-0.055	-0.055	-0.14	-0.09	1	-0.0011
Other Units Involved	0.057	-0.0073	-0.0087	-0.017	-0.034	-0.02	-0.033	-0.086	-0.099	-0.0011	1

7. Correlation Heatmap of Severity and Related Factors

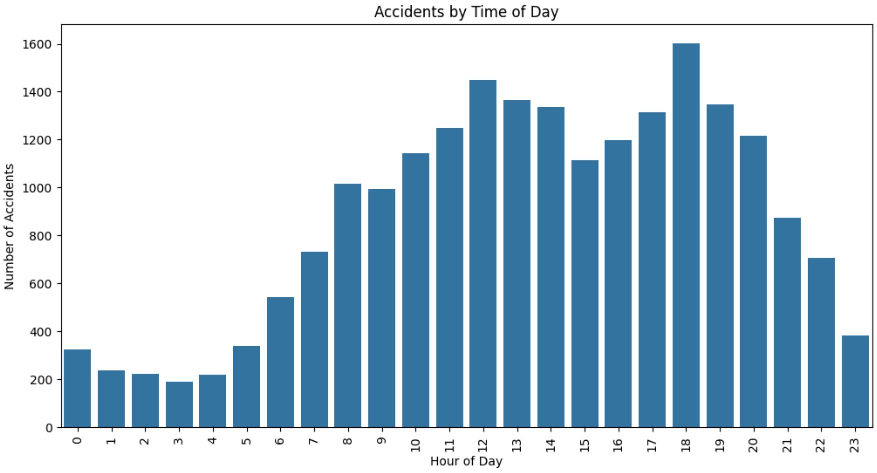
In summary, the visual data suggests a general decrease in both the frequency and severity of traffic accidents in Catalonia over the period from 2010 to 2021, with particularly notable changes occurring in 2020, likely due to the impact of the COVID-19 pandemic on mobility and travel behaviors.

e. On what days of the week and times of day do most accidents occur? Are there notable differences between weekdays and weekends?



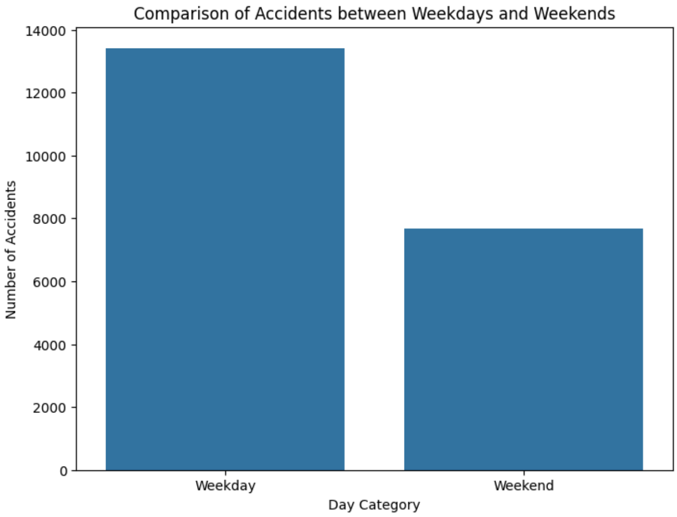
The analysis of traffic accident patterns based on the days of the week and times of day over the given period presents some clear trends. During the weekdays, the occurrence of traffic accidents remains relatively constant, with a minor surge on Fridays, suggesting an association with the end-of-week traffic increase. Conversely, the weekends show a modest decline in accidents, with Sunday recording the lowest figures, possibly reflecting the reduced commuter and business-related travel.

8. Bar chart representing accidents during the week



A closer examination of the daily temporal distribution reveals a significant rise in accidents during the morning, reaching a peak during the midday hours. This pattern aligns with typical morning rush hours when traffic density is high. There is a noticeable dip in the afternoon, followed by secondary peak in the early evening, around 18:00 to 20:00, which can be attributed to the evening commute. As night sets in, the frequency of accidents sharply decreases, corresponding with the reduced volume of vehicles on the road.

9. Bar chart representing accidents by time of day



10. Bar chart representing accidents by day category

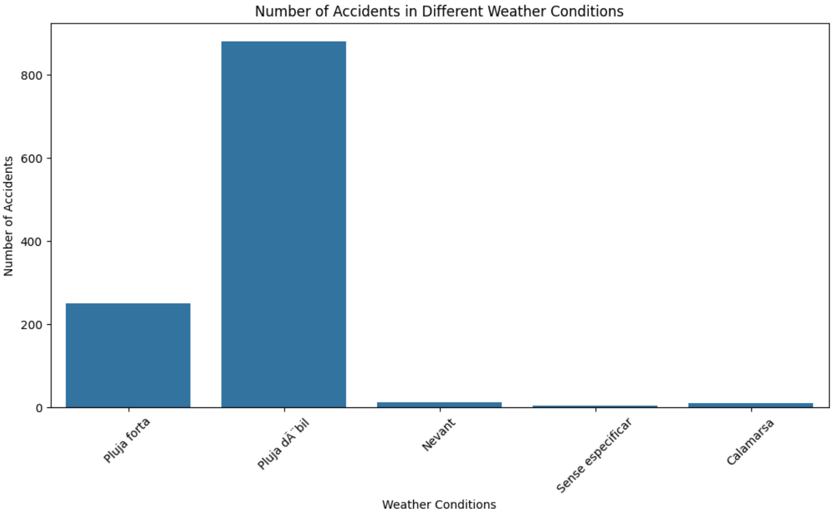
The disparity between weekdays and weekends is further emphasized in the comparative analysis. Weekdays experience a higher incidence of accidents overall, which is in line with the regular patterns of work-related travel and increased road usage.

Weekends, with their more relaxed traffic conditions, see fewer accidents, substantiating the notion that reduced travel and less congested roads can have a positive impact on traffic safety. However, we can notice that in 2 weekend days we have around 8,000 accidents, whereas in 5 weekdays we have just under 13,000. As a result, if road parameters relating to weekends were transferred to weekdays, we would have significantly more accidents.

These findings are critical for informing traffic management policies and safety measures. By targeting high-risk times and adjusting traffic flow controls, it may be possible to alleviate congestion and reduce the potential for accidents, particularly during peak travel hours.

f. How do different weather conditions affect the likelihood of accidents? Is there a correlation between visibility, road conditions, and accident severity?

The visual data below indicates a discernible pattern between weather conditions and the frequency of accidents. To obtain these visualizations, we have selected the part of our dataset where the weather is not good. However, these values are very small, representing almost one-twentieth of our data. The first image shows a bar chart with different weather conditions such as rain and fog, with each bar representing the number of accidents. This suggests a higher incidence of accidents in certain adverse weather conditions, which could be due to reduced visibility, road grip, or driver experience in such conditions.



11. Bar chart representing the correlation between accident and weather conditions

The second and subsequent images (appendix) focus on specific elements such as visibility, lighting, environment, fog, road objects, road surface, and special measures, each of which can impact driving conditions. The bar charts seem to indicate that reduced visibility and poor lighting conditions have a small impact on the number of accidents.

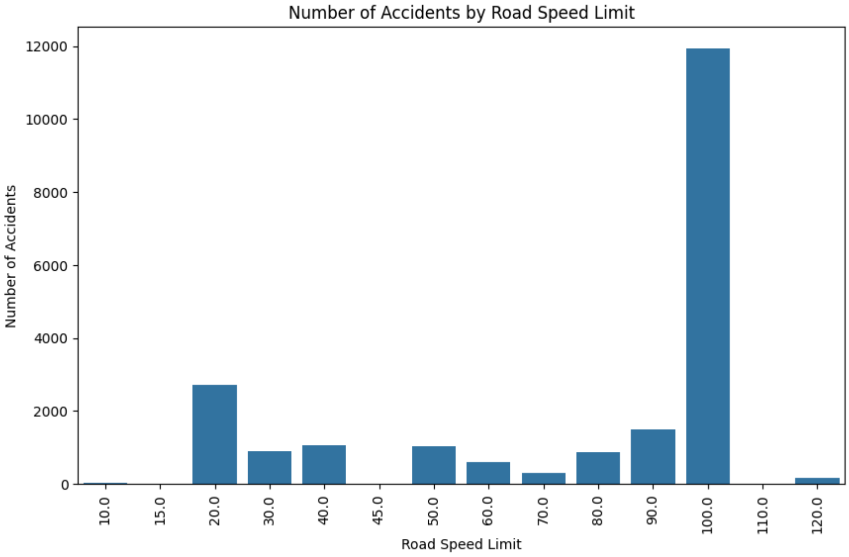
Moreover, the images suggest that certain road environments and the presence of objects on the road do not contribute so much to the probability of accidents. However, the presence of distractions, obstacles or unexpected changes in the road layout for which drivers are not necessarily prepared could theoretically increase the number of accidents. Of course, the number of accidents due to this is not zero, so there's always room for improvement in this area, but it's not very significant.

Finally, the images underline the role of road surface conditions and traffic measures in accident rates. Poor road surface conditions can lead to a lack of vehicle control, while the implementation of special traffic measures might either mitigate or, if poorly managed, contribute to the confusion and subsequent accidents.

In summary, the data visualizations corroborate the hypothesis that weather conditions, along with visibility, road conditions, and other related factors, have a direct impact on the severity and frequency of traffic accidents. The exact nature of these correlations could be further explored with statistical analysis to determine causality and the strength of these relationships.

g. What impact do road features (such as speed limits and road types) and traffic density have on the occurrence of accidents?

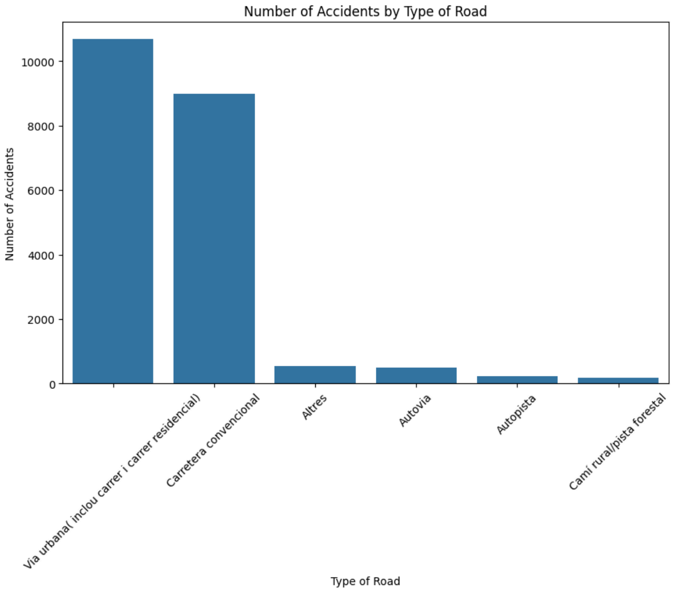
Analyzing the series of data visualizations below, we can construct a comprehensive narrative on the impact of road features, such as speed limits and road types, as well as traffic density on the occurrence of accidents.



12. Bar chart representing accidents by road speed limit

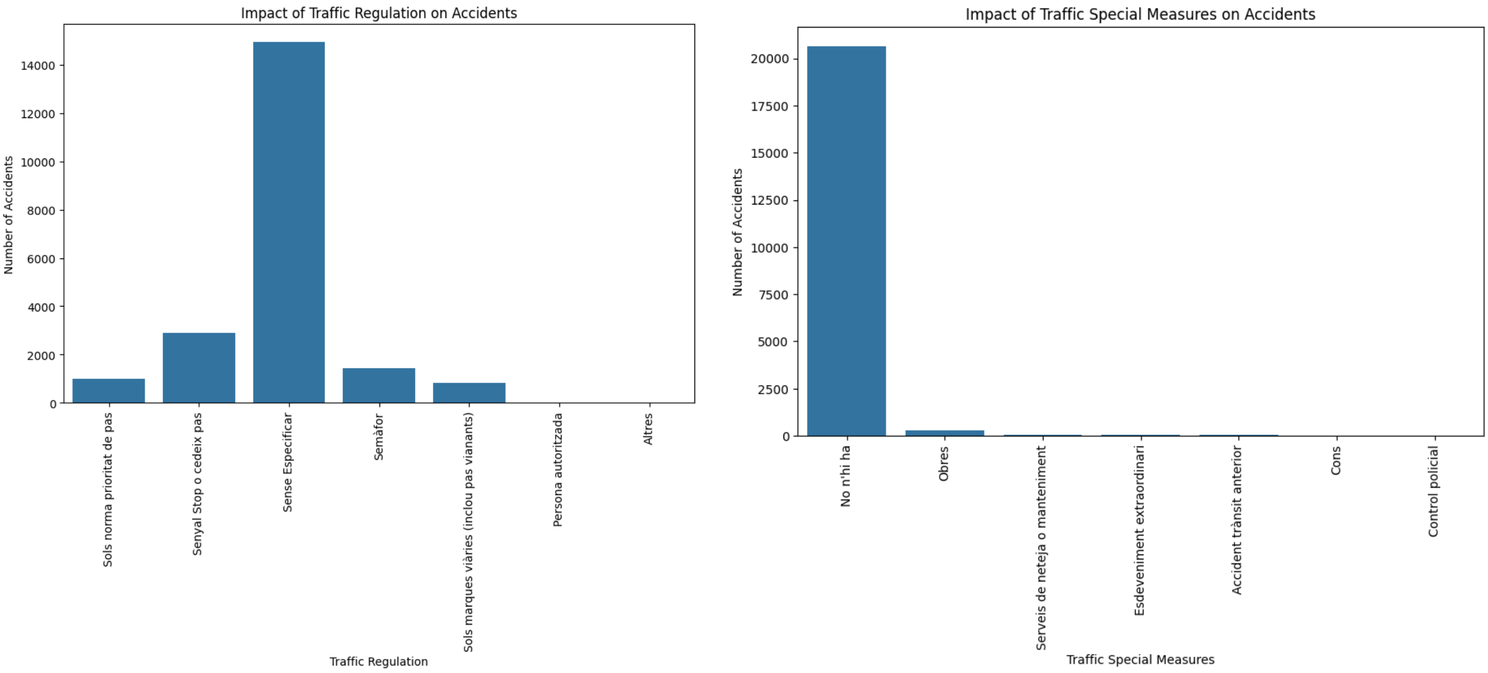
The bar chart titled "Number of Accidents by Road Speed Limit" presents a different trend than initially suggested, with the highest number of accidents occurring at the 100 km/h speed limit. This peak could indicate that roads with this speed limit, which often include rural highways or transition zones between urban and rural areas, are high-risk locations for traffic incidents. These areas may experience a mix of high-speed and local traffic, contributing to a greater likelihood of accidents.

In the chart "Number of Accidents by Type of Road," urban roads ("Via urbana") and conventional roads ("Carretera convencional") continue to show a high number of accidents, which is consistent with the complex driving environment of urban areas. Highways and freeways ("Autovias" and "Autopistas") still report fewer incidents, likely due to their design focused on high-speed travel with safety considerations such as controlled access points and separated lanes.



13. Bar chart representing accidents by type of road

The "Impact of Traffic Regulation on Accidents" chart suggests that areas with traffic signals have a higher incidence of accidents, which could reflect the increased complexity and potential for conflict at intersections controlled by signals. If not, this can underline the lack of respect for regulation and signage. Generally speaking, there are signs in dangerous areas. As a result, there may be a lack of respect for signs and signals. In contrast, areas with fewer regulatory measures like "Give Way" signs or pedestrian crossings without traffic signals show fewer accidents, which may be due to reduced vehicle interactions and more cautious driving behavior.



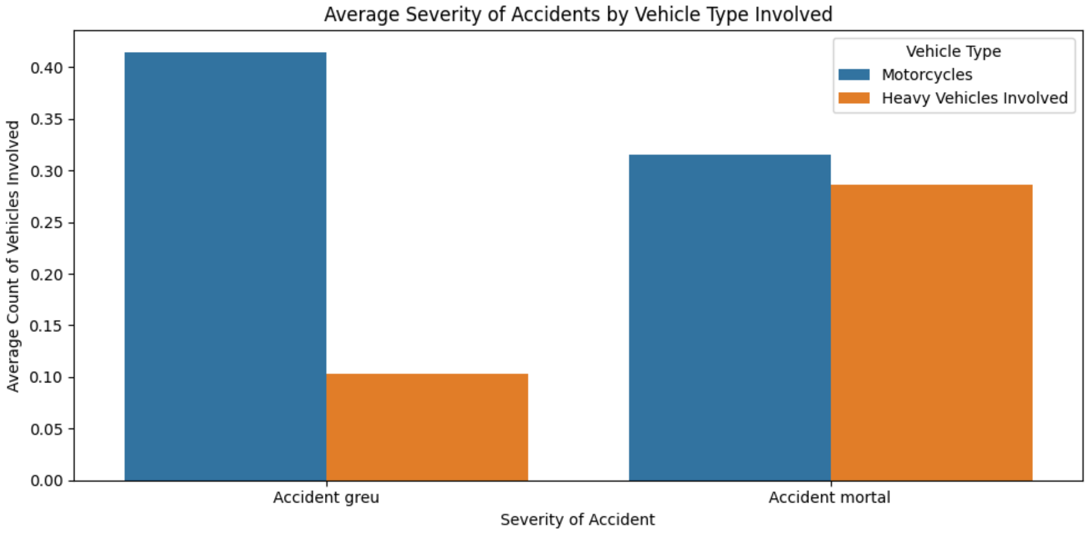
14. Bar chart representing impact of traffic regulation and special measures on accident

Finally, the "Impact of Traffic Special Measures on Accidents" chart illustrates that the absence of special traffic measures is linked to a higher frequency of accidents. This implies that interventions such as roadworks, event-related traffic control, and police checkpoints serve to lower the number of accidents, likely by reducing speeds, improving driver focus, or enhancing road safety conditions.

In summary, these insights suggest that while road features like speed limits and road types influence accident rates, the implementation of traffic regulations and special measures plays a significant role in managing traffic safety. Efficient control and regulation of traffic, especially in areas with higher speed limits or significant traffic signals, can be crucial in reducing the occurrence of accidents.

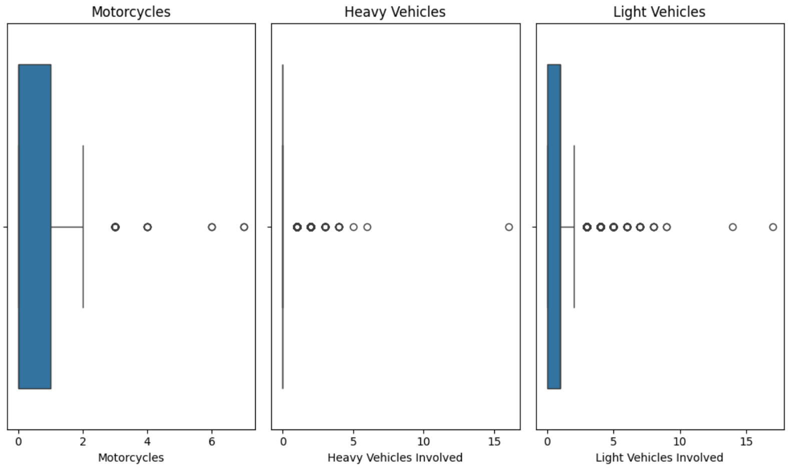
h. Does the involvement of specific types of vehicles (like heavy trucks and motorcycles) correlate with more severe accidents?

The involvement of specific types of vehicles in traffic accidents presents a nuanced picture of road safety. According to the bar chart titled "Average Severity of Accidents by Vehicle Type Involved," motorcycles are less frequently involved in fatal accidents compared to heavy vehicles. However, when accidents do occur, they tend to be of higher severity for motorcycle riders, likely due to the inherent vulnerability of motorcyclists compared to occupants of larger, enclosed vehicles.



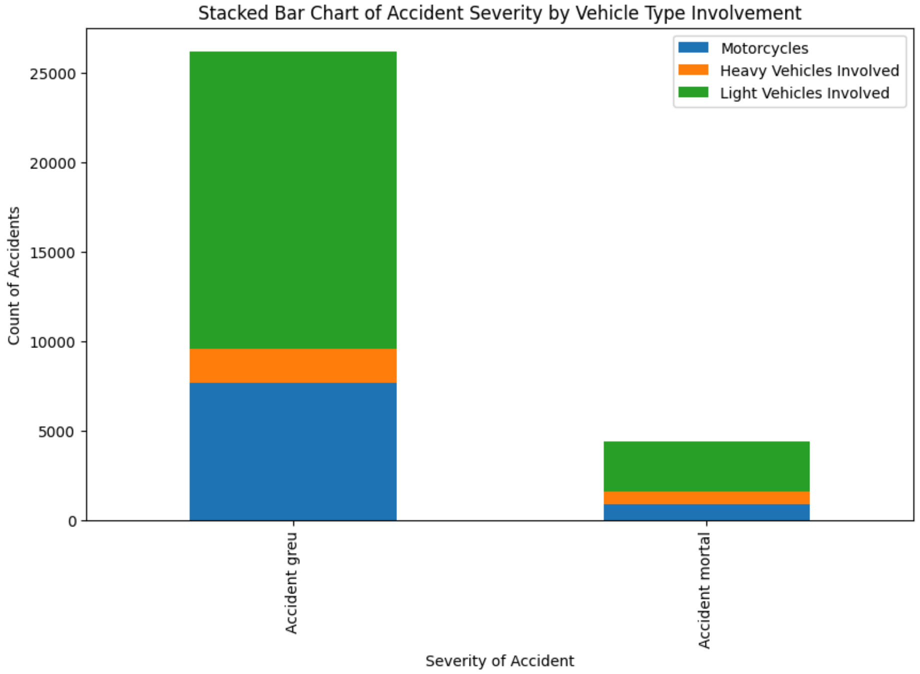
15. Bar chart representing average severity of accidents by vehicle type involved

The distribution of accidents across different vehicle types, as depicted in the stacked bar chart, suggests a higher absolute number of accidents involving light vehicles. This is consistent with their greater numbers on the roads. Heavy vehicles, while involved in a smaller total number of accidents, show a disproportionate involvement in fatal accidents, indicating that when accidents with heavy vehicles occur, they are more likely to result in fatalities.



16. Box plot diagrams representing motorcycles, heavy vehicles, and light vehicles involved in accidents

The box plot diagrams for motorcycles, heavy vehicles, and light vehicles involved in accidents illustrate the spread and outliers in the data. For motorcycles and heavy vehicles, the presence of outliers suggests that while most accidents might result in fewer severe outcomes, there are exceptional cases where the number of vehicles involved in severe accidents is significantly higher.



17. Box plot representing répartition of vehicle type involved in severed accident

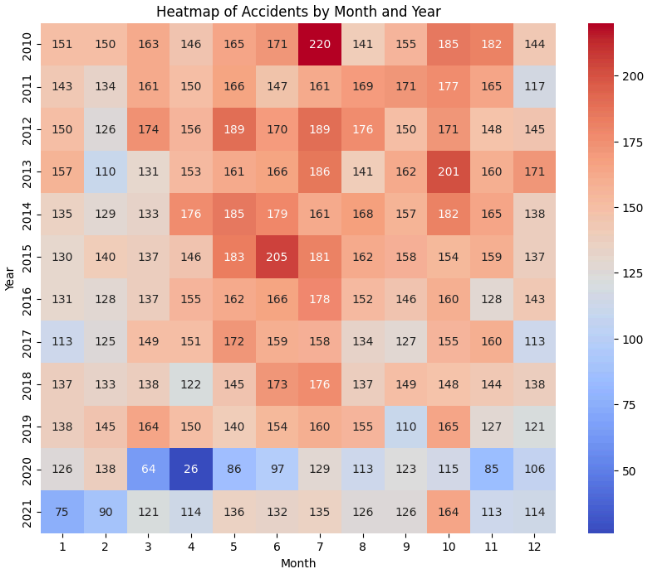
In interpreting these data, it is crucial to consider the overall traffic composition. According to the Bureau of Transportation Statistics, light vehicles, including cars and light trucks, make up the majority of the vehicle fleet and account for the bulk of vehicle miles traveled. In contrast, motorcycles represent a smaller fraction of vehicle miles traveled, and heavy trucks, while essential for freight transport, also contribute a smaller share of total vehicle miles traveled.

The Federal Motor Carrier Safety Administration's Pocket Guide to Large Truck and Bus Statistics indicates that heavy vehicles, such as trucks and buses, are subject to stringent safety regulations due to their potential to cause severe accidents. The data suggest that while heavy vehicles are involved in a lower number of accidents, the severity of these accidents tends to be greater, which aligns with the data visualizations provided.

In conclusion, the proportion of accidents involving motorcycles and heavy vehicles, and their associated severity, must be contextualized within the broader spectrum of road usage. Light vehicles, as the predominant road users, account for the majority of accidents. However, accidents involving motorcycles and heavy vehicles tend to be more severe, underscoring the need for targeted safety measures for these vehicle types. The effectiveness of such measures can be further supported by the FMCSA's ongoing data collection and analysis efforts, which help shape policies aimed at reducing fatalities and injuries on the road.

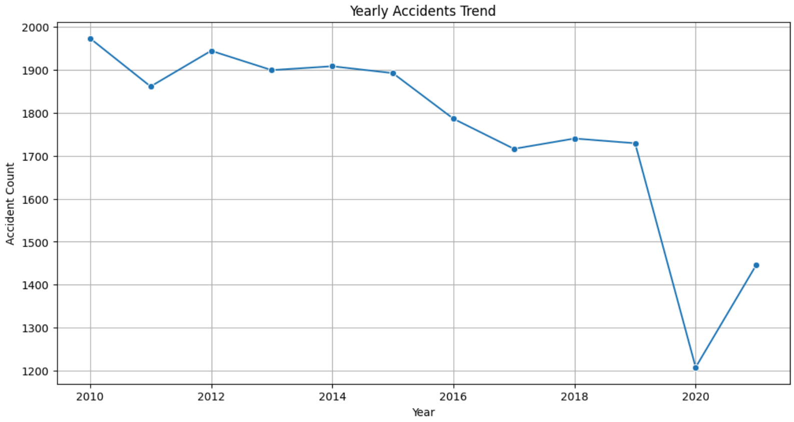
i. Are there specific periods (months, years) where accident patterns cluster significantly? What might be the causes for these clusters?

The heatmap of accidents by month and year displays noticeable variances across different periods. For instance, there are discernible peaks in certain months, such as a pronounced increase in May of 2015 and a significant dip in April of 2020. This could be attributed to various factors including seasonal changes affecting road conditions, increased travel during holiday months, or even the impact of global events such as the COVID-19 pandemic which led to a notable decrease in travel and subsequently, accidents, in April 2020.



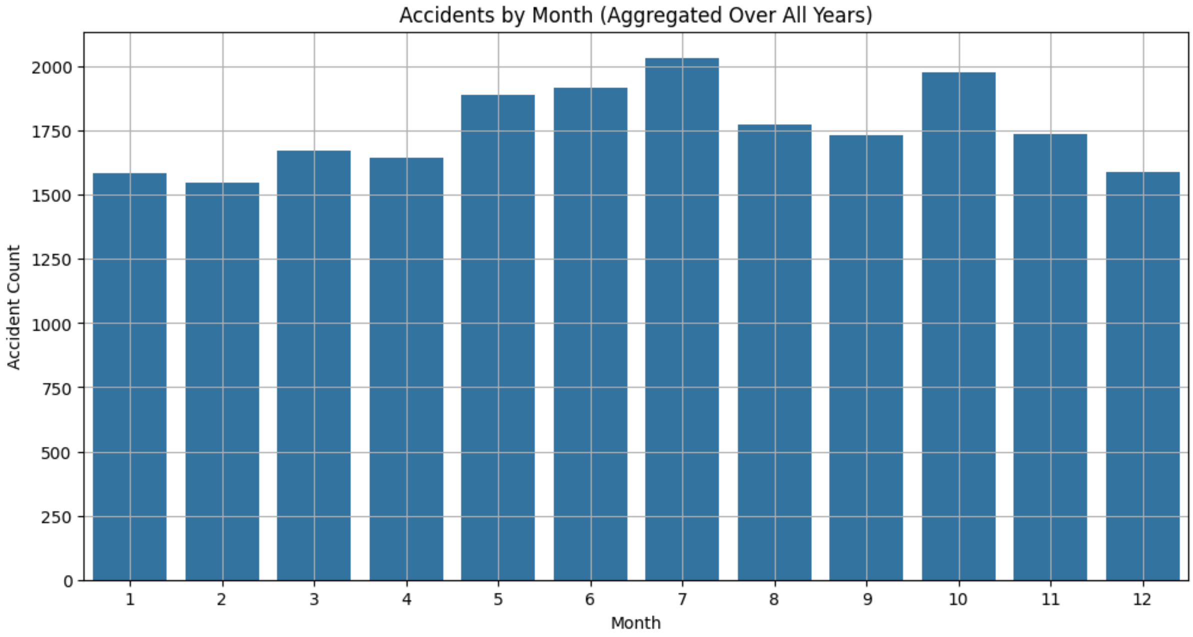
18. Heatmap of accidents by Month and Year

The yearly accidents trend line reveals a general decrease in the number of accidents from 2010, with a sharp decline in 2020, followed by a rebound in 2021. This trend could correlate with broader initiatives for road safety and vehicle technology improvements over the years. However, the steep drop in 2020 and the subsequent increase in 2021 could be influenced by the pandemic-related restrictions and their easing, respectively.



19. Line graph representing yearly accidents trend

The bar chart aggregating accidents by month over all years shows that the distribution of accidents is relatively even, with slight increases in the summer months, potentially due to more active travel and vacationing behavior during this time.



20. Bar plot representing accident by month

Incorporating internet research, it's evident that various factors contribute to these patterns. According to the Insurance Institute for Highway Safety (IIHS), crash fatalities and their rates are influenced by multiple elements such as vehicle types, speed limits, driver behavior, and external factors like weather and light conditions. Moreover, the Federal Highway Administration highlights that crash deaths per miles traveled have fluctuated over the years, with recent increases possibly linked to changes in driving patterns and behaviors.

In summary, the clustering of accidents in certain periods appears to be multifaceted, involving a complex interplay of environmental conditions, societal behaviors, and overarching trends in transportation and mobility. The provided visual data, combined with researched insights, underscores the need for continuous monitoring and targeted interventions to understand and mitigate the risk factors contributing to traffic accidents.

3. Model and Prediction

In the quest to enhance road safety through data-driven insights, this section delves into the development and application of predictive models. Our journey begins with the exploration of traditional statistical models, such as ARIMA, to forecast trends in traffic accidents, fatalities, and serious injuries. Recognizing the limitations of these models in handling complex, multivariate data, we pivot to the realm of deep learning. Here, we introduce advanced models like LSTM (Long Short-Term Memory) networks, renowned for their prowess in deciphering intricate patterns in large datasets. This section not only demonstrates the model building and evaluation process but also reflects on the models efficacy in real-world scenarios. By juxtaposing these methodologies, we aim to unravel the nuanced dynamics of traffic incidents and offer reliable forecasts that could inform policy decisions and preventive strategies in road safety management.

a. Time Series Modeling

Our initial strategy employed ARIMA (AutoRegressive Integrated Moving Average) models, a staple in time series forecasting. This choice was driven by ARIMA's proficiency in capturing trends and patterns in historical data, crucial for predicting future occurrences. The model parameters were meticulously optimized using an automated process (auto_arima), ensuring a robust fit to the data. For example, the ARIMA(0,1,3) model, signifying no autoregression, one-time differencing, and a third-order moving average, was determined as optimal for one aspect of the dataset. This model was adept at rendering the data stationary—a prerequisite for effective ARIMA modeling—and capturing the short-term fluctuations inherent in our data.

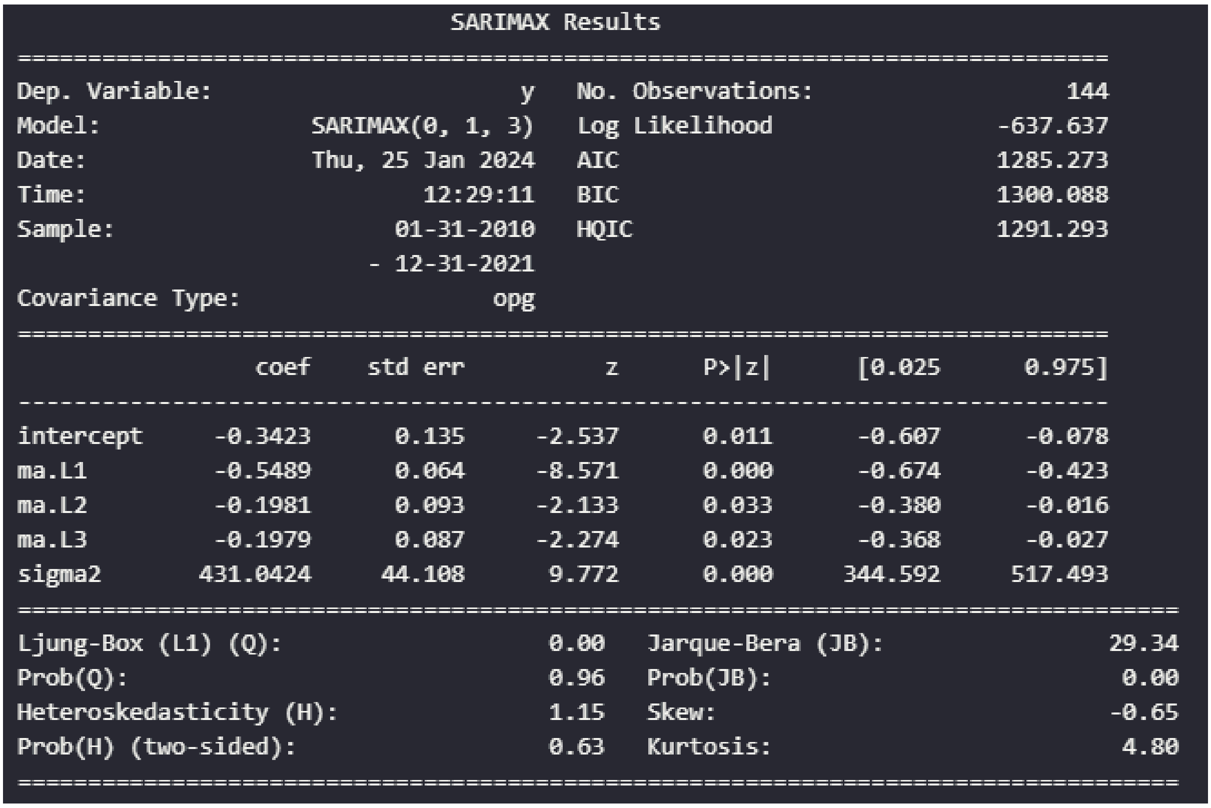
The ARIMA model is a cornerstone in time series analysis due to its flexibility and effectiveness in capturing various data patterns. ARIMA stands for AutoRegressive Integrated Moving Average, combining autoregression, differencing (to make data stationary), and moving average components. This model is adept at analyzing time series data where values are sequentially dependent.

- Autoregression (AR): This captures the relationship between an observation and a specified number of lagged observations (past values).
- Differencing (I): It involves subtracting previous observations to make the time series stationary, addressing trends or seasonality.
- Moving Average (MA): This aspect models the error term as a combination of past error terms.

In our analysis, the ARIMA model was tailored to forecast traffic-related metrics like accidents, fatalities, and serious injuries. The auto_arima function facilitated the identification of optimal parameters, ensuring the model was finely tuned to our specific dataset. For instance, an ARIMA(0,1,3) configuration indicated that the best model did not rely on autoregression, applied one-time differencing for stationarity, and used a third-order moving average.

The model's performance was evaluated based on historical data trends. It demonstrated a significant capability in capturing the immediate past's influence on future values, essential for short-term forecasting. However, its limitation was evident in handling the dataset's multivariate nature, as it predominantly excels in univariate scenarios.

In summary, the ARIMA model played a crucial role in our predictive analysis due to its robustness in dealing with time series data and its ability to provide interpretable results, despite some limitations in handling complex, multivariate relationships.



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  144
Model:               SARIMAX(0, 1, 3)   Log Likelihood                -637.637
Date:                Thu, 25 Jan 2024   AIC                           1285.273
Time:                        12:29:11   BIC                           1300.088
Sample:                    01-31-2010   HQIC                          1291.293
                         - 12-31-2021
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -0.3423      0.135     -2.537      0.011      -0.607      -0.078
ma.L1         -0.5489      0.064     -8.571      0.000      -0.674      -0.423
ma.L2         -0.1981      0.093     -2.133      0.033      -0.380      -0.016
ma.L3         -0.1979      0.087     -2.274      0.023      -0.368      -0.027
sigma2       431.0424     44.108      9.772      0.000     344.592     517.493
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                29.34
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               1.15   Skew:                            -0.65
Prob(H) (two-sided):                  0.63   Kurtosis:                         4.80
===================================================================================
```

21. SARIMAX model summary

The SARIMAX Results from the ARIMA(0,1,3) model provide insightful details about the model's performance and characteristics:

- Model Structure: The SARIMAX(0, 1, 3) indicates no autoregression (AR=0), one differencing (I=1), and a third-order moving average (MA=3). This structure suggests the model focuses on recent errors (moving average part) rather than past values (autoregression part).
- Coefficients: The coefficients for the MA terms (ma.L1, ma.L2, ma.L3) are significant (p-values less than 0.05), indicating their relevance in the model. The negative values suggest an inverse relationship between these terms and the dependent variable.
- Model Fit Metrics:
  - The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) provide measures of the model's quality. Although their absolute values don't tell much alone, lower values are generally preferable when comparing models.

  - The Log Likelihood value indicates the goodness of fit; higher values suggest a better fit.
- Statistical Tests:
  - The Ljung-Box Test checks for autocorrelation in residuals. A high p-value (close to 1) suggests no significant autocorrelation, indicating a good fit.
  - The Jarque-Bera Test assesses the normality of residuals. A low p-value suggests non-normality, which could indicate model inadequacies.
  - Heteroskedasticity test examines the consistency of residual variance across the dataset. A high p-value indicates no heteroskedasticity.
- Additional Observations: The sigma2 value represents the variance of the residuals, giving an idea of the noise in the data.

b. Deep Learning Approach

Given the complexity and multivariate nature of traffic data, we also explored deep learning models, specifically Long Short-Term Memory (LSTM) networks. These models are renowned for their ability to learn long-term dependencies and intricate patterns in large datasets. The LSTM's architecture was carefully crafted, featuring layers that effectively captured the nonlinear relationships within the data. This approach was particularly advantageous for integrating multiple variables that influence the occurrence of traffic incidents.

Our LSTM model was meticulously architected with multiple layers, each designed to capture different aspects and dependencies within the data. This structure was critical in comprehending the intricate relationships between various factors influencing traffic incidents.

The LSTM demonstrated remarkable accuracy, as indicated by its low loss (0.2013) and MAE (0.2006) on the test data. These metrics signify the model's precision in predicting traffic incidents, reflecting its ability to closely align predictions with actual outcomes. Notably, a specific prediction value of 0.9900, particularly in a normalized data context, underscores the model's ability to confidently predict near the higher end of the spectrum.

The success of the LSTM model in our project highlights its effectiveness in managing and interpreting the complexities inherent in multivariate traffic data. It stands as a testament to the potential of deep learning in transforming our approach to forecasting and understanding traffic-related trends and patterns.

c. Evaluation

Both models were rigorously evaluated based on their accuracy, efficiency, and suitability to the dataset's characteristics. The ARIMA model, with its lower complexity, offered a straightforward interpretation but was limited in handling the multivariate aspects of the data. In contrast, the LSTM model, while computationally more intensive, provided a more comprehensive analysis due to its ability to process multiple inputs simultaneously.

The choice of model was influenced by a balance between accuracy and interpretability. The ARIMA model's simplicity made it a reliable choice for univariate forecasting, while the LSTM's advanced capabilities were more suited for capturing the complex interdependencies in multivariate scenarios.

Ultimately, this dual-model approach allowed us to harness the strengths of both traditional statistical methods and modern deep learning techniques, leading to a more nuanced understanding and forecasting of traffic incidents.

Conclusion

In conclusion, our analysis, integrating data visualizations and advanced modeling techniques, offers valuable insights into traffic incident patterns and predictors. As we draw our analysis to a close, it's imperative to reflect on how the insights gleaned can be pragmatically applied. The integration of data visualizations and sophisticated predictive modeling techniques, such as ARIMA and LSTM, offers a comprehensive understanding of traffic incident patterns. These insights are not just academic; they carry significant potential for real-world applications and improvements in road safety. Here are some targeted recommendations based on our findings:

- Enhanced Road Safety Measures: Leveraging the identified patterns and high-risk factors for focused preventive measures in susceptible areas and times.

- Integration with Traffic Management: Utilizing predictive models to inform real-time traffic control systems, adapting dynamically to varying conditions.

- Public Safety Campaigns: Utilizing the data to educate and inform the public about prevalent risks and safe driving practices.

- Collaborative Urban Planning: Employing these insights for informed urban development, emphasizing safer road designs and traffic systems.

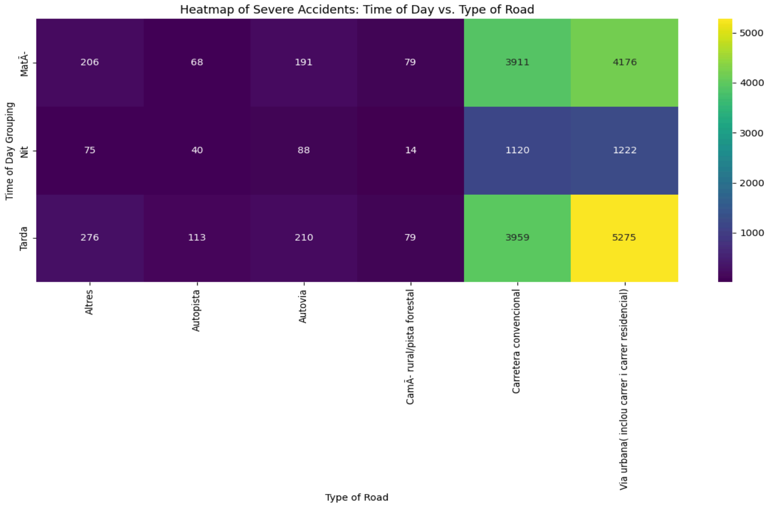
- Ongoing Model Refinement: Continual updating of predictive models with new data to ensure they remain accurate and relevant in changing urban landscapes.

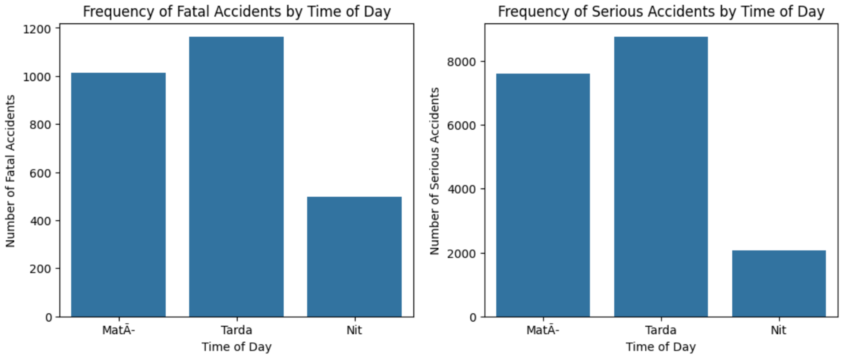
- Expanding Data Integration: Future studies should consider including diverse data sources like social media, smart city sensors, and real-time traffic updates for a more holistic analysis.

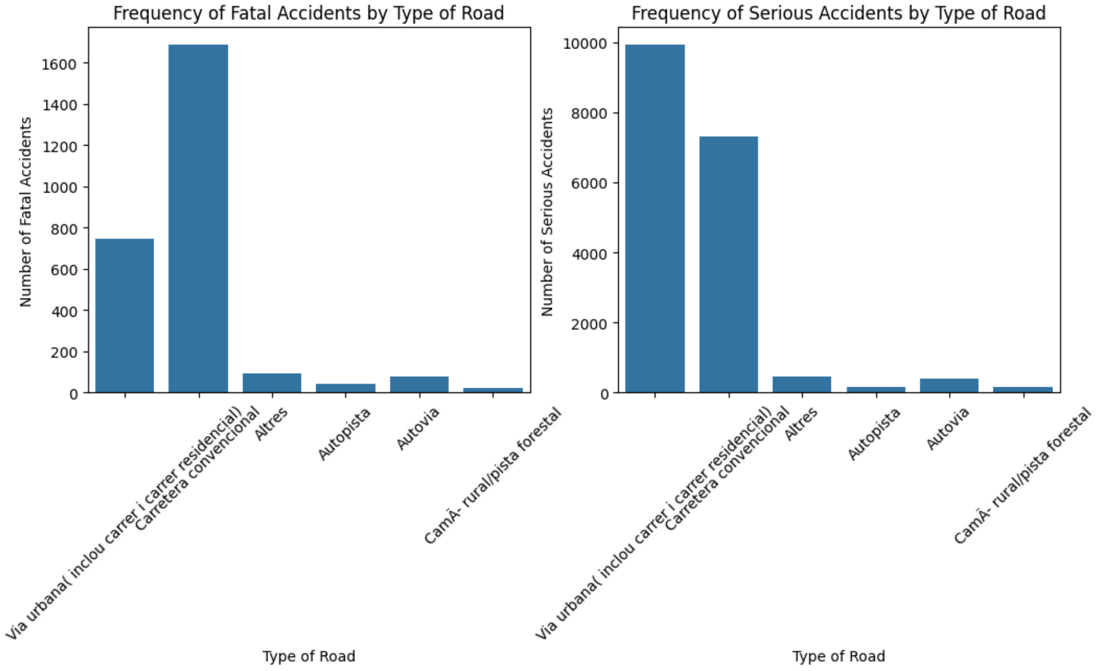
These recommendations aim to harness the power of data and predictive analytics in crafting more effective, data-driven approaches to road safety and urban planning, ultimately contributing to a significant reduction in traffic-related incidents.

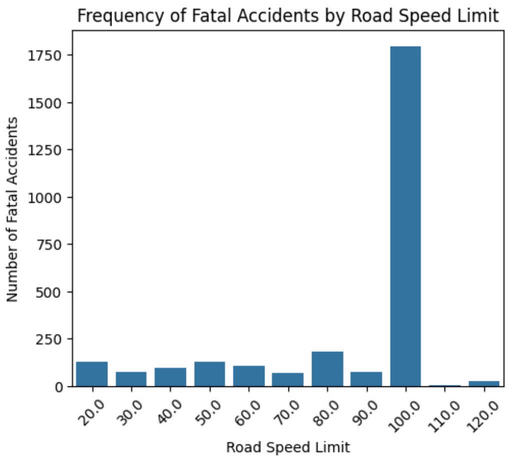
Appendix

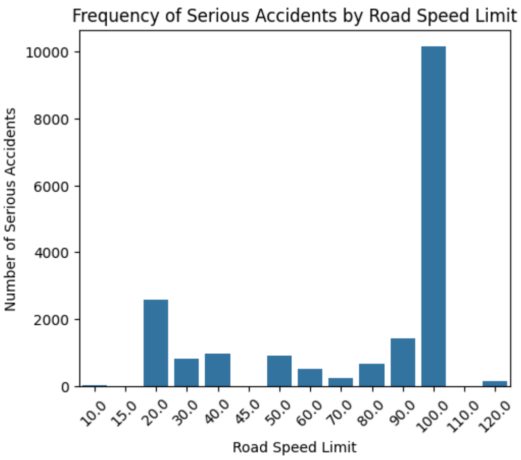
Image related to “What common characteristics (time of day, type of road, etc.) are observed in the most severe accidents?”

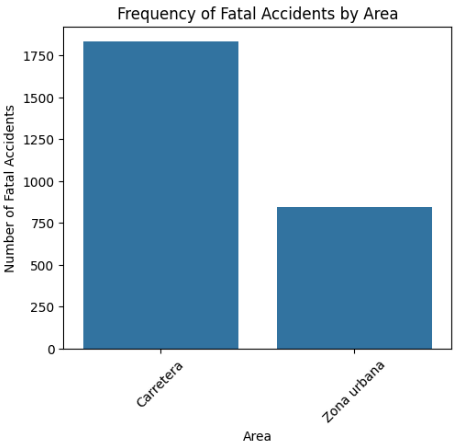


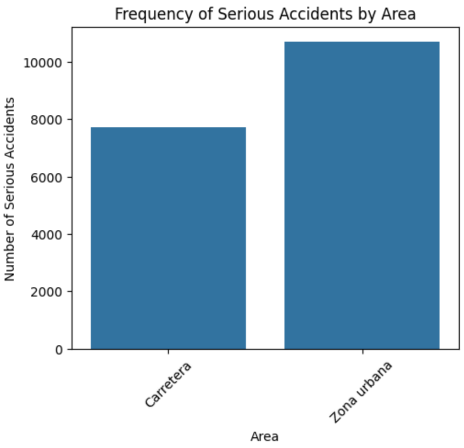


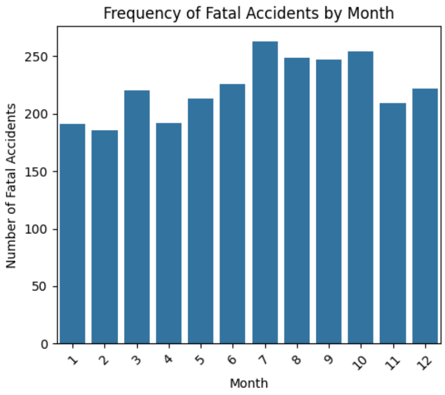


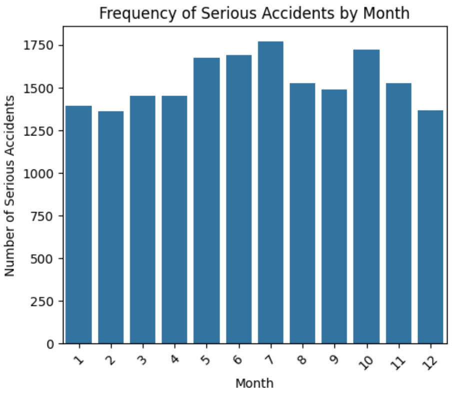

Frequency of Fatal Accidents by Road Speed Limit
Number of Fatal Accidents
0
250
500
750
1000
1250
1500
1750
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
100.0
110.0
120.0
Road Speed Limit

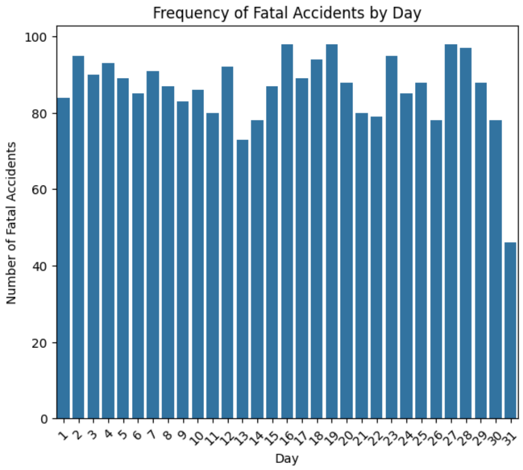

Frequency of Serious Accidents by Road Speed Limit
Number of Serious Accidents
0
2000
4000
6000
8000
10000
10.0
15.0
20.0
30.0
40.0
45.0
50.0
60.0
70.0
80.0
90.0
100.0
110.0
120.0
Road Speed Limit


Frequency of Fatal Accidents by Area
Number of Fatal Accidents
0
250
500
750
1000
1250
1500
1750
Carretera
Zona urbana
Area


Frequency of Serious Accidents by Area
Number of Serious Accidents
0
2000
4000
6000
8000
10000
Carretera
Zona urbana
Area


Frequency of Fatal Accidents by Month
Number of Fatal Accidents
0
50
100
150
200
250
1
2
3
4
5
6
7
8
9
10
11
12
Month


Frequency of Serious Accidents by Month
Number of Serious Accidents
0
250
500
750
1000
1250
1500
1750
1
2
3
4
5
6
7
8
9
10
11
12
Month


Frequency of Fatal Accidents by Day
Number of Fatal Accidents
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
Day

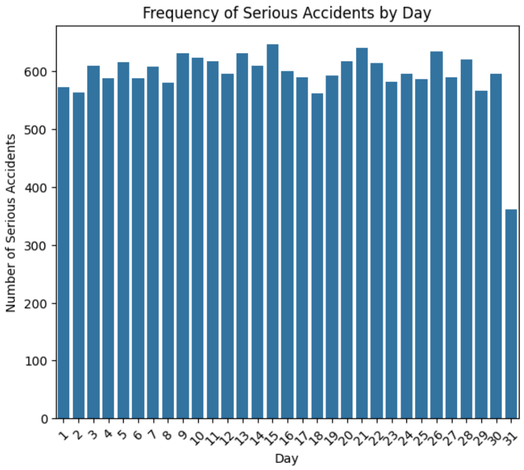
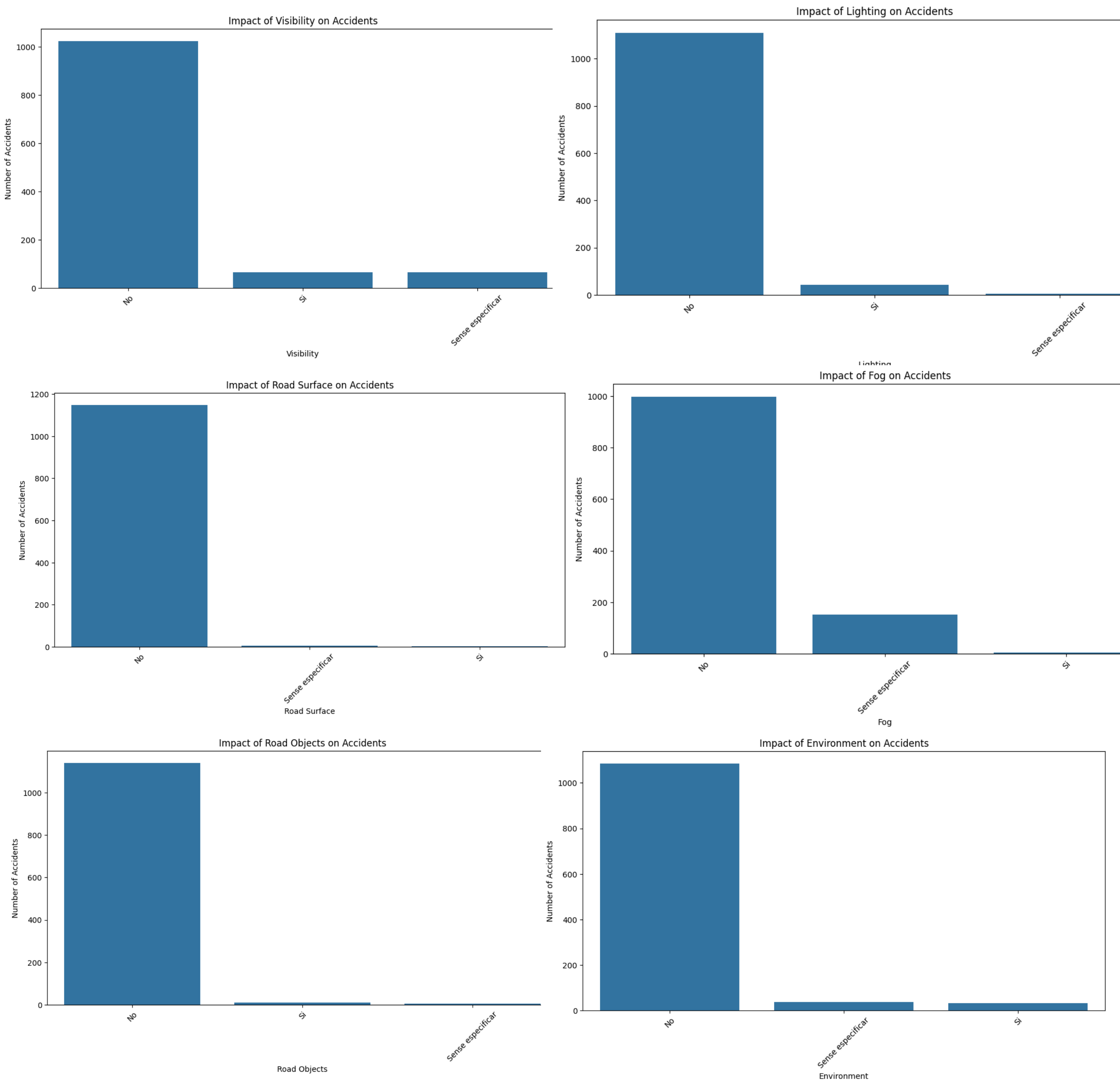
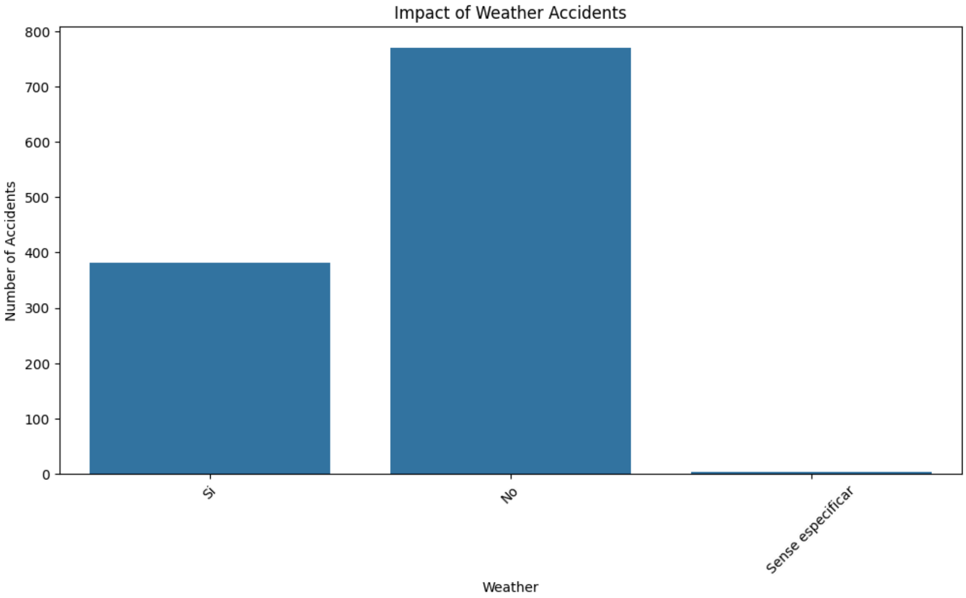
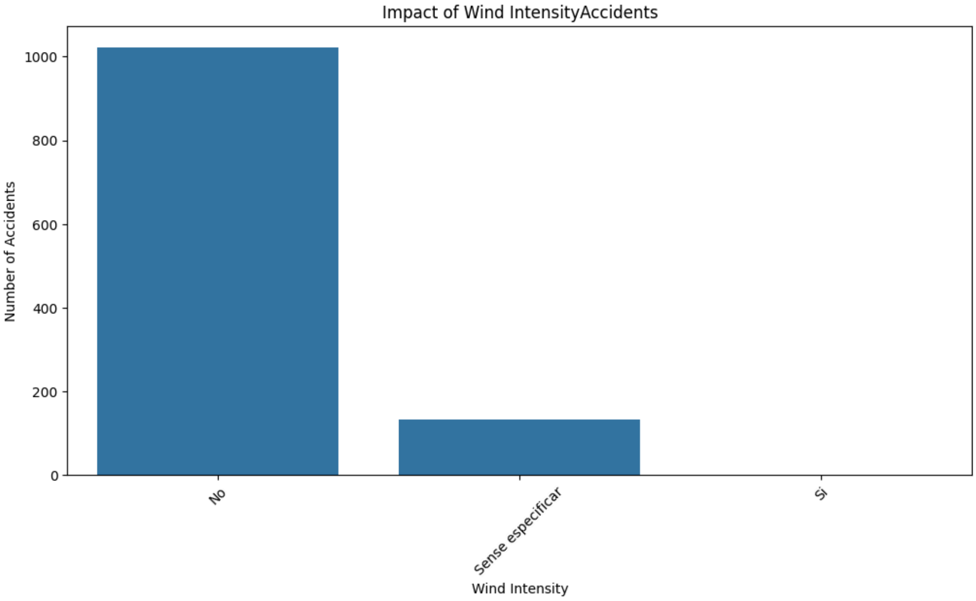
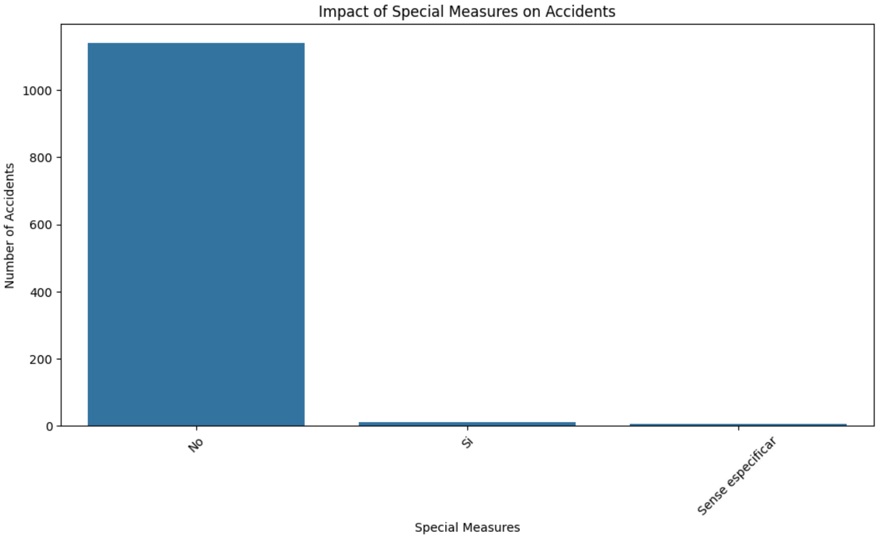

Frequency of Serious Accidents by Day
Number of Serious Accidents
0
100
200
300
400
500
600
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
Day

Image related to “How do different weather conditions affect the likelihood of accidents? Is there a correlation between visibility, road conditions, and accident severity?”




Impact of Weather Accidents
Number of Accidents
0
100
200
300
400
500
600
700
800
Sí
No
Sense especificar
Weather


Impact of Wind IntensityAccidents
Number of Accidents
0
200
400
600
800
1000
No
Sense especificar
Sí
Wind Intensity


Impact of Special Measures on Accidents
Number of Accidents
0
200
400
600
800
1000
No
Sí
Sense especificar
Special Measures

Code

https://github.com/Hugo-SEQUIER/ocean-protocol/tree/main/Traffic%20Accidents%20Catalu%C3%B1a