Understanding Road Accidents in Sardinia: A Data-Driven Approach

Ilaria Coccollone[1]

Abstract

In recent years, road accidents in Sardinia continue to be a topic of contrasts and discussion. The non-decrease in accidents over the years leads to the question of what the main cause is. This research implements a comprehensive data management approach to analyze this phenomenon, focusing on data from ISTAT and OpenStreetMap across different Sardinian provinces. Through a structured pipeline involving data acquisition, integration, quality assessment, and enrichment, were processed and analyzed data about accident components and their locations. The data management process revealed significant challenges in data quality and integration, while enabling the discovery of how infrastructures can significantly influence citizens' driving behavior in everyday life. The findings highlight both the technical aspects of managing road accident data and the persistent impact of this issue on public safety.

University of Milano-Bicocca | MSc Data Science

Contents

Introduction

Sardinia is an island in Italy that brings with it numerous advantages but also significant challenges typical of its geographical isolation from the mainland. One of the most critical issues is the road infrastructure system. Over the years, the island has consistently recorded numerous road accidents, many with fatal outcomes, raising concerns about road safety. This situation particularly affects citizens who rely on continuous mobility between provinces for work, personal commitments, and tourism. This report describes all the phases of the data pipeline designed to study road accidents in Sardinia in the period 2018-2022. In Section 1, the data acquisition techniques are described, focusing on the two main sources: ISTAT and Overpass API. In Section 2, the data integration process is explained, detailing how data from different sources were combined into a single dataset while ensuring uniformity. This section also covers the data enrichment phase, where additional information from other sources was incorporated to improve the dataset and enable more in-depth analyses. In Section 3, the data preparation and data quality phases are described. The dataset undergoes a series of quality checks and preprocessing steps to ensure its reliability before storage. In Section 4, the data storage process is outlined. This includes designing the database schema, selecting the appropriate SQL implementation, loading the data, and optimizing it with additional quality control measures. These steps ensure that the stored data remains complete, consistent, and accurate. Finally, the conclusion highlights the key findings of the analysis, discusses the limitations of the study, and suggests potential future developments.

1. Data Acquisition

1.1 ISTAT Data

The first phase of the project involved the acquisition of data through multiple sources. The primary source was **ISTAT** (Italian National Institute of Statistics), specifically its public use microdata section displayed in Figure 1 concerning road accidents with injuries. Five datasets covering the years 2018-2022 (the most recent available) were downloaded, containing data on all road accidents across the national territory. ISTAT, being Italy's main public statistical research body, ensures high-quality and well-structured data. The CSV datasets were initially filtered in Excel to focus specifically on accidents in Sardinian provinces, retaining only the most relevant columns such as accident location, road type, and number of casualties.

Figure 1: *ISTAT microdata section interface for road accidents data collection*

1.2 OpenStreetMap Data

The second data source was OpenStreetMap (OSM), a free and collaborative global platform for geographic data. Through the OpenStreetMap **Overpass API**, information about key infrastructure in Sardinian territories—such as police stations, schools, and hospitals—was collected using both **requests** and **Overpy** in Python to automate data retrieval and storage in JSON format. For each infrastructure, were extracted information including the unique identifier, geographical coordinates (latitude and longitude), infrastructure name, infrastructure type, address details with street and number, and contact information where available.

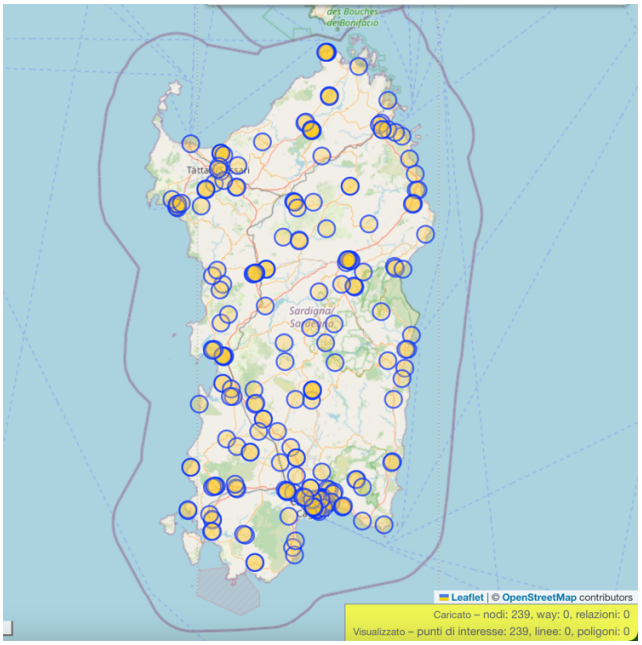

Figure 2: *Geographical Distribution of Infrastructure in Sardinia (Overpass API Visualization)*

1.3 Supplementary Data Sources

To provide additional context, demographic data was acquired from **DEMOISTAT**, specifically focusing on the resident population statistics for each province during the examined period. For 2018, where DEMOISTAT data was unavailable, population statistics were obtained from the **Sardegna Statistiche** website, which provided a comprehensive report on resident population for that calendar year. Additionally, vehicle fleet data for each province and year was collected from ISTAT's public automobile register section, completing our dataset with important contextual information.

2. Data Integration & Enrichment

2.1 Data Integration

After acquiring data from various sources, the goal was to integrate them into a single, clean, and structured dataset. To achieve this, the **Pandas** library in Python was used. First, the pre-cleaned ISTAT datasets on road accidents in Sardinia were loaded into five different dataframes, one for each year from 2018 to 2022. These dataframes were then merged using the pd.concat function in Pandas, resulting in a unified dataset covering the entire analysis period. Next, to integrate data obtained from the **Overpass API**, it was necessary to normalize them into a common format. The JSON file containing information on infrastructure in Sardinian provinces was first converted into a Pandas dataframe. To align this data with the ISTAT dataset, specifically, to assign the number of infrastructures (hospitals, schools, police stations) to each province, a **reverse geocoding** process was used. This was done through the **Nominatim** service from OpenStreetMap, which assigned the correct province based on latitude and longitude. During this process, it was observed that 47 locations returned null values for their provinces. To resolve this issue, an alternative approach was attempted using the **OpenCage Geocoder API**, another free geocoding service that requires an API key. However, this attempt was unsuccessful, most likely due to missing location data in the OpenCage database. Given the relatively low number of affected cases, the missing province names were manually assigned using **Google Maps**, ensuring accurate and complete data integration.

Subsequently, before proceeding with the final integration, the columns containing the name, address, street number, and telephone number of the infrastructures were removed due to an excessive number of missing values. After that, new columns (School, Hospital, and Police) were created to integrate the infrastructure data, counting the number of infrastructures for each province.

Before merging the datasets, the ISTAT dataset contained a column with unique numerical codes for the provinces. This was normalized by assigning the corresponding province name to each code.

Finally, the two datasets were merged using the pd.merge function in Pandas, based on the province column.

2.2 Data Enrichment

To further enrich the dataset through data enrichment, two new columns were added using external data sources.

The first column was the Population data for each province. It was sourced from **DEMOISTAT** for the period 2019-2022 and from **Sardegna Statistiche** for 2018, due to the lack of this year in the primary source. Additionally, information on the number of circulating vehicles per province, in the same period, was obtained from ISTAT Vehicle Data. These additions provide valuable demographic and mobility context, allowing for a more comprehensive analysis of road accidents in Sardinia, considering how the number of inhabitants can increase the movement of vehicles and consequently probably also the road danger.

3. Data Preparation & Quality

3.1 Data Preparation

After integrating the data, several adjustments were made to improve readability and facilitate analysis before storing them in the final dataset. Many qualitative variables were originally represented by numeric codes, making interpretation more difficult. To enhance clarity, these values were converted into their corresponding nominal categories. Variables such as accident location, road type, intersection type, signage, weather conditions, and accident mapping were reformatted as text for better visualization.

Additionally, some variable names were adjusted to more accurately reflect their meaning. For example, the variables related to road classification and intersections were renamed to "tipo_di_carreggiata" and "punto_incidente " to avoid any ambiguity. The year variable was also modified to display the full year instead of a numerical abbreviation, and the counts of infrastructures were converted to integer format for consistency.

During a final quality check, an anomaly was detected in the province column: a unique code that did not correspond to any Sardinian province. This issue was likely caused by an error during data filtering in Excel. To ensure the dataset's accuracy, this incorrect entry was identified and removed before proceeding with the next steps.

3.2 Data Quality

To ensure that the stored data was reliable and accurate, a data quality phase was carried out to assess its completeness, accuracy, and correctness before storage. The first step involved checking for missing values. At first glance, the dataset appeared to have no missing values, which seemed like an excellent result. However, upon closer inspection, it became evident that some columns contained hidden missing data. A more in-depth analysis was conducted by replacing values such as empty strings, tabular spaces, various types of NaN values, and "n.i." (unidentified), revealing that four columns, related to the age of drivers and their outcome (alive or deceased after the accident), contained numerous missing and ambiguous values. Since these columns were both misleading and unhelpful for the analysis, they were removed from the dataset.

After this refinement, an evaluation of dataset completeness confirmed that all remaining columns had 0% missing values. Additionally, checks for duplicate records and outliers were performed. As a measure of semantic accuracy, columns such as the number of deaths and injuries were examined to ensure they contained no negative values. Finally, to verify the correctness of the data, a check was conducted to confirm that the number of deaths within 24 hours was never greater than the number of deaths within 30 days. In cases where inconsistencies were found, the values were corrected accordingly.

4. Data Storage & Analysis

4.1 Data Storage

For Data Storage, a SQL database was chosen for its relational structure that allows to manage complex relationships between road accident factors such as locations, deaths and infrastructures efficiently through primary and external keys, allowing to connect tables while maintaining data integrity through constraints and avoiding inconsistencies in the data and a standardized format. Furthermore, SQL allows to perform complex queries for analysis and facilitates the extraction of specific information and the management of thousands of accidents even if the database wanted to be expanded for future analysis. To create the database, the cleaned CSV dataset generated by the Python script was first imported into a temporary unified table. From this table, four distinct tables were created, each focusing on key aspects of the dataset.

The first and main table, **"incidenti"**, contains detailed information about each accident. It includes a unique accident ID as the primary key, along with columns for the year, province, road type, type of carriageway, accident location, signage, weather conditions, nature of the accident, other vehicles involved, fatalities within 24 hours and 30 days, injuries, time, and quarter.

The second table, **"infrastrutture"**, stores data on infrastructure availability for each province (which serves as the primary key), with separate columns indicating the number of schools, hospitals, and police stations. The third table, **"province"**, includes information on each province along with its population for the respective years. Finally, the **"veicoli_circolanti"** table records the number of registered vehicles in each province for the corresponding years.

4.2 Post-Storage Data Quality

Before running the queries for analysis, a final data quality check was performed to ensure that the data remained reliable after storage, allowing for accurate and meaningful analyses. First, data completeness was re-evaluated by checking for any missing values in each table. Then, semantic accuracy was assessed by verifying the presence of negative values in columns such as deaths and injuries. To ensure consistency, it was rechecked if deaths within 24 hours exceeded those within 30 days and whether province names contained unexpected values. Additionally, temporal consistency was examined by identifying any anomalies in accident counts across different years or quarters, such as an unusually low number of accidents in a specific period, which could indicate missing data or issues in data collection. Finally, regarding the currency metric, it was noted that the dataset includes data up to 2022, which is the most recent year available for public use from the ISTAT website.

4.3 Data Analysis

To cover an analysis on different aspects of the phenomenon of road accidents in Sardinia in the period 2018-2022, a series of SQL queries were performed. The first offers an analysis examining accidents by province considering the total number of accidents, deaths within 24 hours, the number of injured and the fatality rate per accident. as can be seen from the results reported in Table 1 the provinces of Sassari and Cagliari are those with the highest number of accidents, which is in line with the number of inhabitants and urbanization. The provinces of Nuoro and South Sardinia are those with the highest mortality rates, so even though they are provinces with a lower number of accidents, they are those with the highest damage.

Province	Total Accidents	Totals Deaths 24h	Total Injures	Fatality rate per accidents	Fatality rate per injured
Cagliari	4422	93	5896	2.103	1.553
Nuoro	1662	71	2382	4.272	2.895

Oristano	1259	39	1716	3.098	2.222
Sassari	6427	106	9239	1.650	1.134
Sud Sardegna	2316	112	3568	4.836	3.043

Table 1: *General Traffic Accident Statistics by Province*

The second query performed provided a general overview of the distribution of infrastructures both in absolute and relative terms (per 100,000 inhabitants). As reported in Table 2 it emerged that Southern Sardinia has the highest number of hospitals, although considering the availability of the inhabitants Oristano records the highest value. Sassari instead has the highest number of police stations while Cagliari the highest number of schools.

Province	**Total Hospitals**	**Total Police Station**	**Total Schools**	**Hospital per 100k**	**Police per 100k**	**Schools per 100k**
Cagliari	4	19	35	0.950	4.506	8.300
Nuoro	5	27	10	2.495	13.475	5.000
Oristano	6	12	7	3.956	7.912	4.616
Sassari	4	34	29	0.839	7.135	6.086
Sud Sardegna	10	22	15	2.966	6.523	4.450

Table 2: *Absolute and Percentage Values (per 100,000 inhabitants) of Infrastructure by Province*

The third query performed highlighted the distribution of road accidents in relation to weather, another factor that impacts road safety. As reported in Table 3, it emerged that the majority of accidents occur under clear skies. A considerable number of accidents were also detected in conditions of rain and strong wind.

Weather	Clear	Fog	Strong Wind	Rain	Hail	Other
Total Accidents	13800	29	85	1254	29	880
% Accidents	85.837	0.180	0.510	7.780	0.180	5.474

Table 3: *Absolute and Percentage Distribution of Accidents by Weather*

The last query showed the number and percentage of accidents that occurred in the absence of road signs for each Sardinian province. As reported in Table 4, Nuoro has the highest percentage of accidents without signs, followed by South Sardinia and Sassari.

Province	Accidents without Signs	% Accidents
Cagliari	207	4.681
Nuoro	253	15.223
Oristano	123	9.770
Sassari	678	10.549
Sud Sardegna	248	10.708

Table 4: *Absolute and Percentage Distribution of the Number of Accidents by Signs*

Conclusions

In conclusion, this research has highlighted factors that could influence road safety in the various provinces. The results such as those of the province of Nuoro which shows a high mortality and an equally high number of accidents without signaling lead to reflect on the safety of road infrastructures. Other factors such as weather conditions, number of schools or number of hospitals and number of police stations could influence the speed of response and the speed of rescue. However, this analysis has limitations, the lack of data on speed, type of vehicles or age of the driver could omit important details on the dynamics of road accidents. For this reason, for future analyses, it would be possible to study these aspects more deeply and analyze more how certain factors can safeguard or on the

contrary endanger the lives of citizens, in a region, with an evident problem of mobility and road infrastructures.

References

ISTAT. (2024). Demo ISTAT. Istituto Nazionale di Statistica.
https://demo.istat.it/app/?l=it&a=2019&i=POS

ISTAT. (2025). Esplora Dati. Istituto Nazionale di Statistica.
https://esploradati.istat.it/databrowser/#/it/dw/categories

ISTAT. (2024). Rilevazione degli incidenti stradali con lesioni a persone. Istituto Nazionale di Statistica.
https://www.istat.it/microdati/rilevazione-degli-incidenti-stradali-con-lesioni-a-persone-3/

ISTAT. (2023). Veicoli e Incidenti Stradali. Istituto Nazionale di Statistica.
http://dati.istat.it/Index.aspx?DataSetCode=DCIS_VEICOLIPRA#

Sardegna Statistiche. (2018).
https://www.sardegnastatistiche.it/documenti/12_103_20181212133014.pdf

Ondata. (2025). Guida API ISTAT.
https://ondata.github.io/guida-api-istat/

OpenCage. (2025). Geocoding Dashboard.
https://opencagedata.com/dashboard#geocoding

OpenStreetMap Contributors. (2025). Overpass API.
https://overpass-api.de

YouTube. (2020).
https://youtu.be/0OfsXybrweI