Data Integration & Visualization

by

Siro Brotón, Ivan Ocheretianyi, Anastasiia Sadova, Filip Markovic

NIA: 100496683, 100487317, 100494620, 100575101

Contents

1 Motivation

For this project, we have decided to provide a data source showing unbiased statistics for national and regional crime, immigration and population statistics. We aim to solely collect data from independent sources.

Crimes and immigration are both very important political topics in the current time in Europe. With this current issue, we are seeing a strong increase in European far right parties as reported by NBC news or POLITICO. This issue is highly polarizing, dividing the political spectrum. Left leaning parties raise fears over right wing parties gaining government power and increasing problems for migrants rather than solving them, while right leaning governments blame crime rates on immigration and foreign nationals (DW news)

Our goal is to collect data that can be analyzed to spot correlations between crime and immigration rates in countries. This allows users to the political narrative for sociological theories such as the Immigrant Revitalization Theory, which suggests that immigrant populations can stabilize and reduce crime in communities.

Although there do exist many reports on that topic, independent and unbiased research further proves important, as both sides of the political spectrum work with biases when compiling reports. Our goal is to provide a clean data source that allows further research on this topic.

This project describes a technical overview of the data integration process involving the extraction, transformation and loading of data; known as an ETL process. In the end we show a proof of concept that implements a small part of this project leaving room for scalability, to show that our project idea can also work on a bigger scale.

2 Data sources

The initial phase of this project is to identify and select the data sources we will use. Since the project's scope aims to cover the entire world , rather than a single city or country, several different approaches to data acquisition must be considered.

2.1 Regional data

The first data acquisition strategy prioritizes sub-national databases from specific regions and cities. This approach provides a higher level of data granularity, allowing for the analysis of location-specific details such as the precise neighborhood where an offense occurred and the inclusion of valuable micro-data often omitted from national-level aggregations.

2.1.1 Madrid

To illustrate this sub-national data acquisition strategy, we present a case study of the city of Madrid. The primary resource identified is the open data portal of the Madrid City Council (*Portal de datos abiertos del Ayuntamiento de Madrid*). The portal features a search engine, which we queried using the Spanish term for crimes, *delitos*.

This query returns a variety of granular datasets, including statistics on crimes by location, detailed crime rates, perceived citizen insecurity by neighborhood or district, police patrol figures, and offenses committed on EMT Madrid buses, among other relevant data points. This high-resolution data is invaluable as it provides a level of detail often lost in national-level aggregations. Furthermore, a subsequent query for *Policía Municipal* (Municipal Police) grants access to more specific law enforcement datasets.

We examined one such dataset in detail: Statistical data on Municipal Police actions. This data is available in .xlsx format and contains monthly statistics on police interventions. The contents include:

- Actions related to Citizen Security by district.
- Individuals arrested and investigated by offense type (e.g., injuries, domestic and gender-based violence, abuse of children, threats, sexual abuse, and theft).
- Arrest and investigation figures broken down by district.
- Reports on animal sales, hunting, and fishing.

This level of detail allows for a more targeted analysis. For example, a preliminary review of this data indicates that the districts with the highest number of arrests or investigations are **Center** and **Puente de Vallecas** (1,229 and 911 people, respectively).

Finally, we also explored the main Madrid City Council website. While it also features a search tool, we observed that it ultimately redirects to the open data portal, reinforcing the portal's role as the city's central repository for public data.

2.1.2 London

Similarly, London serves as another prime example of a city with a mature open data ecosystem. The London Datastore, the official, free open data portal for the city, is managed by the Greater London Authority. Analogous to Madrid's portal, it provides a vast range of datasets on topics such as air quality, housing, transportation, and demographics.

Of particular relevance to this project is the MPS Monthly Crime Dashboard Data. This dataset is published by the Metropolitan Police Service (MPS) and is reliably updated on the sixth day of each month. It provides granular data on knife crimes, "other crimes," and "total notifiable offences," all of which are available for download in .xlsx and .csv formats, making it highly suitable for data ingestion and analysis.

2.2 National data

An alternative methodology involves a "top-down" approach, focusing on national-level data rather than sub-national sources. This strategy significantly reduces the number of data sources required.

A primary advantage of this method is that national data is typically already pre-harmonized, providing a consistent data structure for the entire country. This obviates the complex task of integrating and standardizing disparate datasets from various cities.

For this approach, data is sourced from the official national statistics institute of the respective country (e.g., Spain's INE, Germany's Destatis). These institutions aggregate and publish comprehensive datasets from numerous national sources, serving as a single, authoritative repository.

2.2.1 Spain

For instance, the national-level source for Spain is the INE (*Instituto Nacional de Estadística* or National Statistics Institute). Data from such official institutes is typically well-structured, multi-dimensional (allowing for filtering), and internally consistent, significantly streamlining the data processing phase.

While the INE covers a vast range of topics (from economy to demographics), its societal statistics are directly relevant to this project. We can find detailed Statistics on Convicted Persons for both Adults and Minors. These datasets offer granular breakdowns by year, type of crime, and various demographic characteristics such as sex, age, and nationality.

As an alternative or supplementary national source, the official open data portal for the Government of Spain (*Portal de datos abiertos del Gobierno de España*) can be used. This platform serves as the central repository where all public bodies of the Spanish government—from national ministries to regional agencies—publish their datasets for public access. As an official government initiative, it provides authoritative, first-party data directly from the source.

2.2.2 France

Following the national-level approach, France serves as another key example. The French Republic maintains an official, centralized portal, data.gouv.fr. This platform is operated by the Interministerial Directorate for Digital (DINUM), which confirms its status as an authoritative and reliable first-party source.

A key feature of this platform is its dual-access model, providing data through both downloadable datasets and a public API. The portal covers nearly every sector of public administration, with extensive data on Economy and Business, Environment and Energy, Transport and Mobility, Health, Education and Research, and Demographics.

For the specific scope of this project, we can find highly relevant datasets, such as the Municipal, departmental and regional statistical databases of delinquency recorded by the national police and gendarmerie. This data is available in multiple formats suitable for analysis, such as .csv and .xlsx, as well as in document formats like .pdf and .docx.

2.3 Supranational and EU Institutions

While sourcing data from national institutes is more efficient than a granular, city-by-city approach, it introduces its own significant challenges. The first is the language barrier, as national portals are often available only in the local language, complicating data discovery and interpretation. The second is the challenge of scale; managing separate data acquisition pipelines for all 44 countries in Europe is a substantial undertaking.

This is where **supranational sources** provide a clear advantage. These platforms aggregate data from multiple countries, or even the entire continent, into a single, harmonized repository. Critically, this data is provided in English and other international languages, which removes the language barrier and is a definitive benefit for a pan-European analysis.

2.3.1 European Union

Eurostat serves as the official statistical office of the European Union (EU) and plays a pivotal role in ensuring the availability of reliable, comparable, and harmonized statistical information across all EU Member States. Its primary function is to collect, standardize, and disseminate statistical data that reflect the economic, demographic, and social realities of the European Union, thereby supporting evidence-based policymaking and fostering transparency within the European governance framework.

Importantly, Eurostat does not conduct primary data collection itself. Instead, it functions as a central coordination and harmonization body, aggregating data produced and validated by the national statistical institutes of each Member State—such as Spain's *Instituto Nacional de Estadística* (INE), Germany's *Statistisches Bundesamt* (Destatis), or France's *Institut National de la Statistique et des Études Économiques* (INSEE). This decentralized but standardized approach ensures the methodological consistency of European statistics and allows for meaningful cross-country comparisons.

In addition to EU Member States, Eurostat's datasets also encompass statistical information from selected non-EU countries, particularly those in the European neighbourhood and Mediterranean region. These include, among others, Algeria, Egypt, Libya, Israel, Moldova, Georgia, and Ukraine—reflecting Eurostat's broader engagement in regional and international statistical cooperation.

Eurostat provides comprehensive datasets across a wide array of thematic domains, including: General and regional statistics, Economy and finance, Population and social conditions, Industry, trade and services, Agriculture, forestry and fisheries, International trade, Transport, Environment and energy, as well as Science, technology and the digital society. Within these thematic areas, users can access high-quality data on diverse topics such as population dynamics and criminal statistics. For instance, relevant datasets include *Police-recorded offences by offence category* and *Population by age group, sex and country of birth*.

To facilitate access and usability, Eurostat's data are made available through multiple formats. Researchers, policymakers, and the general public can download datasets directly in commonly used formats such as CSV, TSV, or SDMX, or access them programmatically via Eurostat's publicly available API. This infrastructure supports a high degree of interoperability, enabling users to conduct complex quantitative analyses and integrate Eurostat's data into broader research workflows and digital applications.

2.3.2 United Nations Organization

The United Nations (UN) is a global intergovernmental organization with the articulated mission of maintaining international peace and security, to develop friendly relations among states, to promote international cooperation, and to serve as a centre for harmonizing the actions of states in achieving those goals.

And such a big organization has its own open data portal. It can be used as general search, but it is better to use the following ones for different particular purposes:

- UNODC (United Nations Office on Drugs and Crime) - Direct and specialized global crime data.
- The World Bank Open Data - Global socio-economic and development data.
- WHO's Global Health Observatory
- UNICEF Data - Global data on the situation of children and women.
- UNESCO Institute for Statistics - Global data on education, science, culture, and communication.

A lot of data can be found there for all countries, for example Persons convicted or Arms seizures. Data can be downloaded in spreadsheet format.

2.3.3 INTERPOL

INTERPOL, The International Criminal Police Organization. This is an intergovernmental organization with 196 member countries.

The databases are for global police cooperation. They allow police in member countries to share and check information on criminals and crimes instantly. It contains sensitive law enforcement information, including:

- Names of criminals and wanted persons
- Fingerprints and DNA profiles
- Stolen and lost travel documents (passports, visas)
- Stolen vehicles
- Information on weapons and other threats

The main issue is that the database is private. However, anyone can apply to become an authorized user of some databases, using special form. Nevertheless, user can obtain valuable information from their publicly available reports and publications including statistical trends, threat assessments, operational summaries and analysis and insights.

2.4 Open and Third-Party Sources

An alternative data acquisition strategy involves leveraging **third-party data providers**. These are often commercial entities that specialize in aggregating large volumes of information from diverse public and proprietary sources.

This approach can be highly efficient, as the provider has already performed the complex work of data collection and, in many cases, harmonization. By utilizing such a service, our workflow would be significantly simplified. Instead of manually sourcing data from numerous national portals, we could simply query the third-party platform for a specific city or crime to retrieve the pre-processed data directly.

2.4.1 EpData

EpData is a platform of the Europa Press news agency that collects, analyzes, and offers public data, especially in graphic and database formats, to facilitate understanding of topics such as the economy, population, employment, and crime, providing context and facilitating the verification of figures.

We can use it as another data source for small cities like Getafe or the whole communities like Madrid or Andalucia. The data can be downloaded in .json or .csv files, so it is easy to work with them.

2.4.2 Statista

Statista is a major online platform for market and consumer data. It is a huge, privately-owned database that gathers statistics and studies from thousands of sources and presents them in an easy-to-use format like charts and infographics. It is run by a German company called Statista GmbH. It's a commercial business, not a government or intergovernmental organization.

It covers an enormous range of topics across 170 industries. User can find data on everything from the market share of smartphone brands to consumer opinions on sustainability. It aggregates information from market research reports, trade publications, scientific journals, and government sources. The topics user can particularly find there: Consumer Goods and FMCG, E-Commerce, Economy and Politics, Energy and Environment, Internet, Technology and Telecommunications, Transportation and Logistics, Travel, Tourism and Hospitality

For a particular instance, we can obtain Most common crimes in El Salvador from June 2023 to May 2024. The information is available in various formats to suit different needs: a raw dataset (.xlsx), a full report (.pdf), a standalone chart (.png), and a presentation slide (.ppt). The only problem that arises is that some of the data is only available with subscription or special academic account.

2.4.3 CrimeoMeter

CrimeoMeter is a commercial platform that provides crime data as a service, primarily through an API. It offers real-time crime data, crime maps, and statistical information for over 700 cities, mostly in the US but with some global coverage. It is designed for businesses and developers to integrate into their own products (e.g., real estate websites, safety apps, news platforms). It is operated by a private company, CityCop Corporation and data used is from various police departments on crime incidents, 911 calls, and sex offender registries.

CrimeoMeter provides data primarily in JSON format.

This is standard for an API-first service. When you make a request to one of their APIs (like the Crime Data API or Sex Offenders API), the server will send back the crime data structured as a JSON object.

For their visual products, the format is different:

- Embeddable Crime Map: it is delivered as an HTML snippet that the user could embed on a webpage or as a raw crime incident data (locations, types, dates) in JSON fromat
- Crime Map: it is delivered as an image file format, such as PNG or JPEG or Geospatial Vector Format (like GeoJSON or KML):

2.5 Automated Data Extraction Using Web Crawlers

In some cases, certain relevant datasets may not be directly available through standardized APIs or open data portals. For example, smaller municipalities, regional police departments, or research institutions may publish statistics only through their websites in tabular or textual format. To bridge this gap, an additional data acquisition method can be implemented through automated web crawlers.

A web crawler is an automated program designed to systematically navigate web pages, extract structured information, and store it in a usable format. This approach allows us to extend the scope of our data collection beyond pre-existing databases and public APIs, ensuring that users can access updated and comprehensive information even when such data is not included in our internal repository.

The crawler-based acquisition process typically involves the following steps:

1. Identification of relevant online data sources that provide publicly accessible statistical information.
2. Extraction of relevant tables, metadata, and textual data using parsing techniques (e.g., HTML, XML, or JSON parsing).
3. Normalization of the extracted content to ensure consistency with the project's existing schema and structure.
4. Validation and filtering of collected data to ensure compliance with data quality standards and legal constraints.

This method is particularly useful for supplementing official data sources, as it allows for the retrieval of information that may be published irregularly or not exposed through machine-readable interfaces. For example, if a specific country's open data portal does not provide an API for crime statistics, a crawler can automatically collect the same information directly from the website's published tables.

To maintain transparency and ethical standards, all crawler implementations should comply with the websites' rules and data usage policies. The extracted data should be limited to publicly available content and used exclusively for research and statistical purposes.

Integrating web crawlers into the data acquisition pipeline thus provides a flexible and scalable solution for real-time data retrieval, ensuring that the platform remains dynamic and up-to-date even in cases where no standardized API or downloadable dataset exists.

3 Data processing

In order to proceed working with the data correctly and further analyze it, we need to normalize and clean it, ensuring future consistency and adequacy of the results.

By gathering and balancing data from multiple sources, we can guarantee that every record follows the same structure and meaning, which is essential when comparing information from different regions or institutions.

Proper data processing not only improves reliability and accuracy, but also lays the foundation for meaningful statistical analysis and visualization in the later stages of the project.

This phase will involve the following tasks:

- Filtering, cleansing, removal of duplicates, validation and authenticating of data.
- Performing calculations, translations, or summarizations based on the raw data. Examples: converting currencies or other units of measurement, editing text strings, and more.
- Removing, encrypting, or protecting data (data anonymisation).
- Formatting the data into tables or joined tables to match the schema of the target data warehouse.

3.1 Data preprocessing. Filtering, cleansing, removing duplicates, validating and authenticating the data

3.1.1 Normalization

- Numerical values

In this project we are working on a data that is already mainly normalized. Despite the fact that our main fields of interest are described with a natural numbers, some of them can be represented in the different manners. Let's take as an example the way we write numbers of the 10^6 range. It can be represented in a next four ways:

- 1 000 000
- 1000000
- 1.000.000

In this work we will follow the basic representation without any separators inside of the number (second example from the above).

It is also worth taking into an account the way we will represent floating point numbers throughout the analysis. All floating-point values will be rounded to one decimal place, with the decimal part separated from the whole number by a point (ex. 8.64 will be rounded to 8.6; 8.57 will be rounded to 8.6).

- Date and time

Since we will be working with the statistical data related to the years and the months, it is important to normalize their representation. As it was mentioned previously, in the proof of concept we will be working exclusively on the yearly data due to its small scope. Despite that, we will cover here transformations to be applied to the full-scaled project.

The year will be expressed as a four-digit number, starting from 2018 to the 2022, in order to avoid gaps caused by missing data from earlier periods. Months will be represented as natural numbers ranging from 1 to 12.

Any other representations, such as categorical {jan, feb, etc..} or {january, february, etc..} will be converted into this numerical format for consistency.

It is also important to note that, in some cases, data may be aggregated by quartiles or triples. When the field of interest is monthly data, the value for a given month will be calculated as the average of the corresponding quartile or triple. This ensures that the data remains consistent and meaningful across the dataset.

- Categorical values

All categorical fields, such as country, nationality, or type of crime should be converted to the same format/case across all the data resources we are working with to ensure consistency. In order to conduct smooth analysis, all the text values will be converted to the lowercase, removing extra spaces. All the acronyms (such as United Kingdom, UK, or Great Britain) we will be uniformly mapped according to a standardized naming table to prevent mismatches and ensure reliable data integration. We will also ensure that different data providers group regions in the same manner, as we can find sources that group Wales, Ireland, Great Britain... into UK and other that don't make such a fine distinction.

- Language normalization

We will talk about it further in the subsection 3.2.1.

- Column naming consistency

Since we will be working with multiple data sources, a mismatch in between attributes may occur. To avoid losing essential data, we will implement standard naming convention for all of the attributes. This will allow us to merge data tables without any conflicts arising and essential data being lost, keeping its integrity. For example:

- Countries will be normalized to a standard lowercase long form, with all variations (ISO codes, abbreviations) mapped to a single consistent representation.
- Numeric values will be standardized with floating-point numbers being rounded to one decimal place with a point as a decimal separator (.), ensuring consistency across datasets.
- Data validation will be applied to remove invalid or missing entries before merging.

All the attributes names should be represented in the snake case format (lowercase letters separated by underscores) to ensure uniformity across all datasets.

3.1.2 Cleansing and filtering

In addition to data normalization, we will perform data cleansing and filtering. In order to optimize storage and resources used, entries that are irrelevant to the workflow or constant values that do not contribute to the analysis will be removed from the working database. Furthermore, we will verify the integrity and validity of all remaining records to ensure that values fall within realistic and expected ranges. Any inconsistencies, missing entries, or duplicate records obtained when merging data will be identified and corrected or discarded when necessary.

3.1.3 Validation and authentication

Data validation and authentication is going to consist of checking whether values lie in the realistic and useful range, as well as ensuring that all data fields correspond to the expected format and type. As an example for that, country's population numbers can not be negative number. Additionally, categorical data such as country names will be cross-checked against standardized lists to avoid any mismatches.

Data validation and authentication will focus on ensuring correctness, consistency, and usability of the data. Each field will be checked to ensure it matches the expected type, format, and range. For example:

- Population figures must be non-negative integers. This check is done just by comparing population value with a suitable range.
- Yearly data must follow chronological order. It is already achieved due to the database initial structure and the way we merge/upload data into it.
- In the bigger scale projects, it might come in handy to cross-check the categorical attributes, such as country names for example, with reference tables to prevent mismatches (e.g., ”United States”, ”USA”, and ”US” all mapped to the same standard name). It is done using library pycountry in python, that contains full country’s names and their possible abbreviations. Countries’ names are being converted into lowercase and stripped, after what they are getting compared with the library’s entries and normalized if needed.
- Calculated metrics, such as crime per 100,000 inhabitants or immigration rates, will be validated for plausible ranges and consistent rounding. This rate is going to be rounded to 2 decimals.

This step ensures that the dataset is not only accurate but also ready for integration, analysis, and reporting, supporting reproducible and trustworthy results.

3.2 Performing calculations, translations, or summarizations based on the raw data. Examples: converting currencies or other units of measurement, editing text strings, and more.

3.2.1 Translation

A significant challenge in aggregating data from regional and national sources is the linguistic heterogeneity of the non-numerical data. To integrate these sources, a robust strategy for both translation and semantic harmonization is required.

A preliminary approach, such as direct machine translation, is insufficient. This method fails to address the more complex issue of semantic inconsistency. Even within a single language, the same concept may be represented by numerous synonyms (e.g., ”people convicted” could be variously labeled as ”convicted”, ”people convicted”, ”condemned”, or ”sentenced”). Attempting to unify tables with such discrepancies would result in fragmented data and failed integration.

To resolve this, we can establish a master schema that uses English as the standard. All data from other sources must be mapped to this schema. This mapping can be operationalized in two ways:

- **Programmatic Mapping:** This involves developing comprehensive lookup tables or a thesaurus to map all known synonyms to the standard term. More advanced implementations could utilize Natural Language Processing models to compare the semantic similarity of fields, automating the matching process and reducing the need for manual pre-translation.
- **Manual Curation:** The ”gold standard” for accuracy would be to engage professional translators, ideally subject-matter experts, to manually translate and map the data files. This approach would ensure the highest data fidelity by correctly interpreting legal and cultural nuances in terminology. However, the primary drawback is that it is a resource-intensive solution, incurring significant financial costs that may not be feasible for the project’s scale.

3.2.2 Calculations and summarization

Despite having described previously the translation and structural transformations, several calculation and summarization steps were applied to standardize and prepare the datasets for analysis

- Immigration rate: in order to ensure consistency in the cross-country comparison, we have decided to ensure that immigration rate is scaled to be represented per 100.000 inhabitants.
- Average metrics: since our main goal is to conduct a meaningful analysis, we need to have metrics calculated accordingly to represent our interest. In order to do so, we will be computing yearly averages for the immigration and crime rate (ex. average crime rate per country committed by a certain group).
- Summarization: after finishing all of the data processing described above, it is important to remember that all of the data we were working on should be merged into the final dataset. In order for that to go smoothly, all datasets are aligned by `country_iso3_id` and `year_id`, allowing direct joins in the database and ensuring that any analysis or visualization reflects clean and standardized information. This merged dataset serves as the foundation for loading into the database, enabling consistent analysis across countries and years.

3.3 Conducting audits to ensure data quality and compliance

3.3.1 Accuracy

The sources that we are using are either official data sources like the UN, City Councils, National institutes of statistics; or other licensed sources that get data from (or from other licensed sources like police), so our data should be the most accurate that can possibly be found online.

To guarantee that during translation no problem arises, professional translators would be hired to translate all names needed. And also, to avoid accidental deletion of data, the comparison of the initial data file and the one translated will be performed using simple comparison script. This would guarantee that no data are missing, no errors are allowed, and thus the user would be able to get accurate data with no modification that would reflect original source.

3.3.2 Completeness

In some cases, some data fields can be empty. As an example, the UN data about people convicted has almost no information about Ukraine or Republic of Moldova. In such case there are a couple of options: we can use other data sources to fill it in, in this case we can use national sources because they are very reliable. In other cases, some data cannot be taken from different data sources as they may not have it, so the other option would be just to leave it blank in case no other alternative exist.

As an alternative option, we can use AI to check if data from another source matches (by the name of field and table) the one that is missing, and if so, insert it. Sometimes it would not even be necessary to use AI would as all the regional data would already be translated and structured, thus simple program would easily handle taking data from one table to insert it into another table.

So the approach we would use is mixed. To avoid some big gaps, for example, in UN data, we would insert regional data with the use of the kind of program described before. This approach would be used for important data like population, migration rate, people convicted that were taken from supranational sources. However, in case of national sources or regional ones, the best approach would be to leave it empty as it would be either impossible or very complicated to fill in data.

3.3.3 Data Protection and Privacy

Our approach to data protection and privacy is built on two core activities: verification and transparent compliance.

First, we will conduct audits to verify that all data sources are fully anonymized and contain no personal information. While our sources are public and should already be anonym as the sources are mostly governmental, this verification step is a crucial part of our due diligence to ensure we comply with data protection regulations. This can be achieved by the script that would check if any personal information (name, surname, passport) are in data sets.

Second, to ensure legal compliance with the terms of use for every dataset, we will create and maintain an internal license log. This log will detail the provider, source URL, and specific license requirements for each dataset. This internal process will inform our public-facing ”Data Sources” page, where we will provide users with a transparent list of our sources and their terms, fulfilling all attribution requirements.

3.3.4 Ethical Use

We recognize that crime statistics are sensitive and can be misinterpreted or weaponized if presented without care. To ensure our project is used responsibly, we are not including any overall report that could have been made by a biased person, we solely include pure statistics and let user compare it relevantly with socio-economic factors like income levels or unemployment rates wherever possible. This helps prevent simplistic conclusions and provides a more nuanced understanding of the data.

3.4 Removing, encrypting, or protecting data (data anonymization)

This subsection is mostly related to the big project and not to the proof of concept that we will cover, but it is still very important and worth mentioning when working with big scale projects.

This step ensures that no personal or sensitive information is used or exposed during the analysis. Since the project relies on open-source and publicly available datasets, this process mainly involves verifying that no personal data or corrupted entries appear within the collected resources. In case such information is detected, we will take responsibility for removing it entirely or, if necessary, encrypting it using the AES encryption algorithm with a 256-bit key, which offers a strong balance between security and performance and is well suited for protecting files or database fields.

When storing the data, we would also ensure that all of the potentially sensitive files or database entries are protected with restricted access. These measures guarantee compliance with privacy standards and help maintain the ethical and responsible use of data throughout the project.

4 Storage

In this project, the main goal is the creation of a secure and efficient data source for inspecting crime data and combining it with immigration data. To achieve this, all datasets will be stored in a relational PostgreSQL database, which ensures data integrity, structured organization and reliable access control.

4.1 Database decision

We formulated the following requirements for the design and implementation of our database system to ensure that it supports the analytical objectives of the project while maintaining robustness, scalability and long-term maintainability:

- The database must provide native and efficient support for structured data, as all input datasets possess a well-defined schema and consistent attribute structure. This enables predictable querying, facilitates schema validation, and ensures compatibility across multiple data sources.
- It must include robust mechanisms for representing and managing relationships among entities, since the data is inherently relational in nature. Key entities such as countries, population counts, crime statistics, immigration rates, and temporal indicators are closely interconnected, and preserving these relationships is essential for ensuring analytical accuracy and referential integrity.
- The system should strictly enforce data integrity and consistency by applying declarative constraints and referential rules derived from a carefully constructed and implemented Entity–Relationship (ER) model. This design enforces the logical structure of the data, prevents redundancy and anomalies, and supports seamless integration of future data extensions.
- Given the analytical scope of the project, a relational SQL-based database is preferred due to its strong adherence to ACID (Atomicity, Consistency, Isolation, Durability) principles, mature transaction handling, and proven performance in managing relational datasets. Additionally, our team's prior experience and familiarity with SQL environments allow for efficient schema design, optimization, and maintenance of data pipelines.
- Finally, the database should be designed with scalability and security in mind, allowing for the integration of additional datasets and user access levels in future project stages. Implementing appropriate indexing, normalization and access control mechanisms will ensure both high query performance and data confidentiality.

Due to the given reasons, we concluded that a relational database would be the best choice. We decided on PostgreSQl as our database management system, since our team has most experience with it specifically and, if needed, the high extensibility of the database allows us to use extensions and implement complex rules and queries.

4.2 National vs. Regional data

In the majority of our sources national crime and immigration statistics are published yearly, while regional data is published monthly. As our goal is to solely provide a long term observation of crime and immigration, we decided to only save the national crime statistics per year and the regional crime statistics per month.

Although monthly data is more detailed, the integration of these different sources makes the implementation more difficult, since we need to find data sources serving monthly data. This would be outside the scope of this project and hurt our choice in data sources.

For this reason our project will use two databases. One for complete national data and one for incomplete regional data.

4.3 Database draft

Figure 1 shows how national data could be saved into one single table. It contains differentiation of criminals per type of person, and all defined data for the construction of our database.

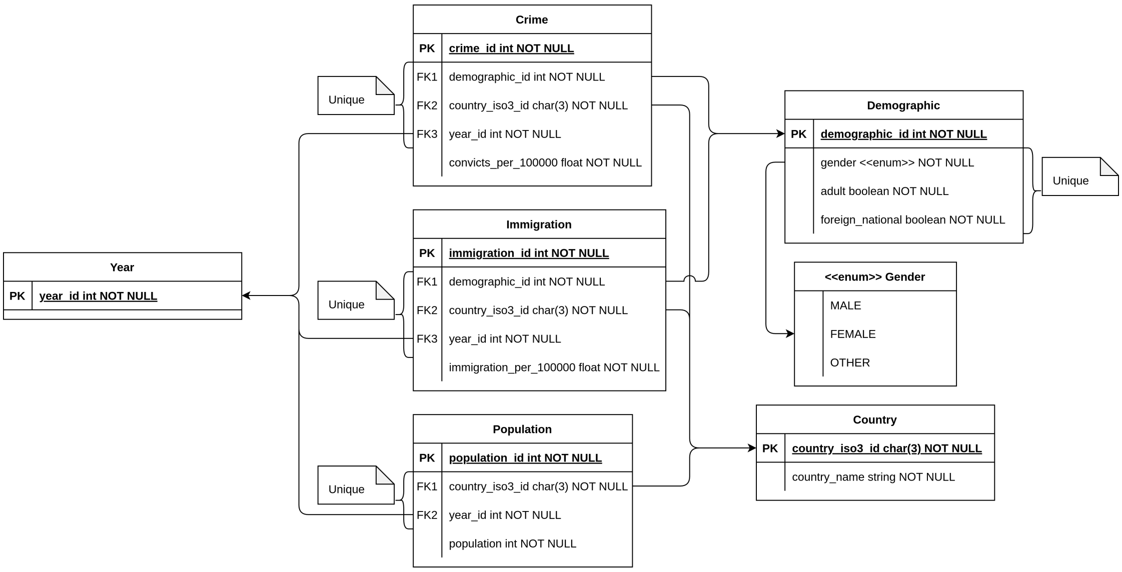
Crime
PK crime_id int NOT NULL
FK1 demographic_id int NOT NULL
FK2 country_iso3_id char(3) NOT NULL
FK3 year_id int NOT NULL
convicts_per_100000 float NOT NULL
Unique
Immigration
PK immigration_id int NOT NULL
FK1 demographic_id int NOT NULL
FK2 country_iso3_id char(3) NOT NULL
FK3 year_id int NOT NULL
immigration_per_100000 float NOT NULL
Unique
Population
PK population_id int NOT NULL
FK1 country_iso3_id char(3) NOT NULL
FK2 year_id int NOT NULL
population int NOT NULL
Unique
Year
PK year_id int NOT NULL
Demographic
PK demographic_id int NOT NULL
gender <<enum>> NOT NULL
adult boolean NOT NULL
foreign_national boolean NOT NULL
Unique
<<enum>> Gender
MALE
FEMALE
OTHER
Country
PK country_iso3_id char(3) NOT NULL
country_name string NOT NULL

Figure 1: ER diagram draft

A similar database draft can be designed for the regional database. The regional database, however, has to handle missing values better, since how we earlier defined, getting data for every single region worldwide proves difficult.

4.4 Database updates

While we designed an ETL process for our data sources, the structure of these data sources can always change. This means that the data processing part of our project is subject to change yearly, which can mean a significant amount of work. We therefore decided that the updating of the database should be done yearly.

Yearly, we will use manual and automatic checks to inspect if data sources changed. After inspection and potential fixes in our code the process of loading data into the database is the same as before.

After these updates on the databases on our server, our new database will contain data for the new year.

5 Infrastructure

Figure 2 provides a high-level view of our architecture. The data sources listed are described in the chapter Data providers. Data Processing will be implemented with Python, as described in the chapter Data processing. The database used is PostgreSQL as described in the chapter Storage.

The application will run on a server where we will upload our downloaded data, run our data processing scripts and save the results in our database; corresponding to the previously defined ETL process.

6 About the proof of concept

The goal of our proof of concept is to show our presented project can work as we described it in a real life scenario. The main objective of this project is to provide a database with national data for crime and immigration, which we realized in our proof of concept. The only added complexity of a full implementation of this project compared to our proof

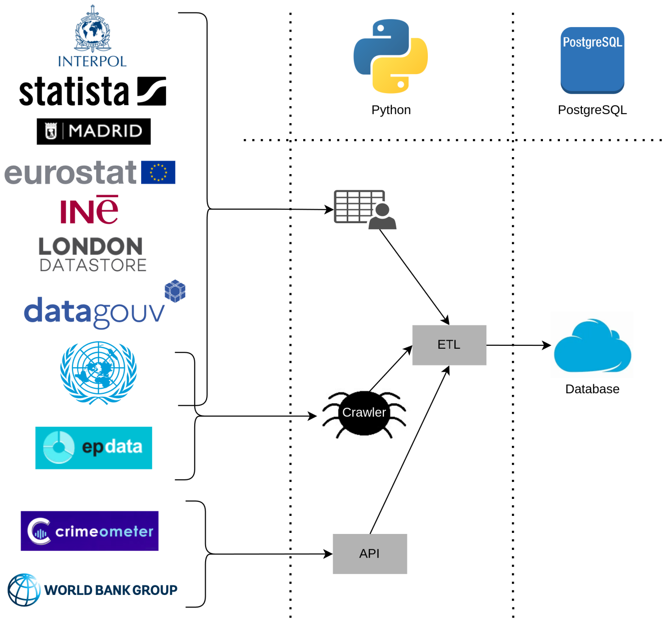

Figure 2: Infrastructure diagram

of concept would comes from adding data sources and huge quantities of data in the processing.

The whole implementation of the proof of concept, and a copy of this document can all be found in our GitHub repository.

Our proof of concept will focus solely on European countries to avoid too large a scope.

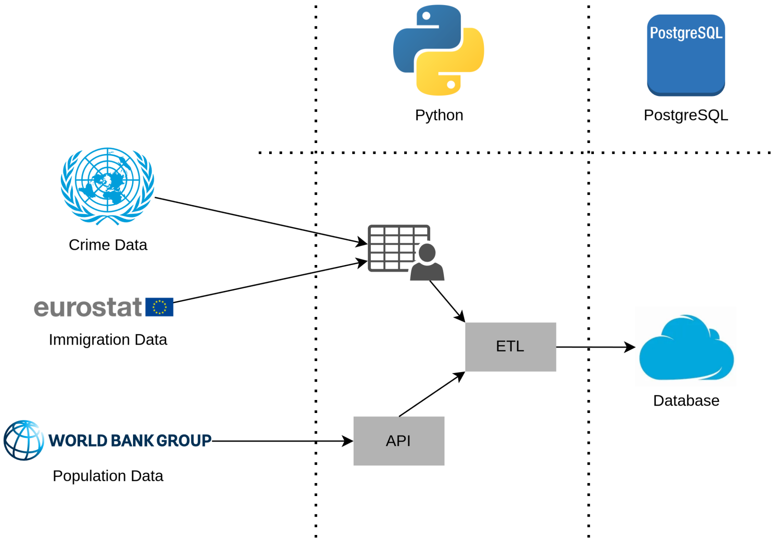

Figure 3: Proof of concept infrastructure diagram

Figure 3 provides a high-level view of our architecture.

6.1 Data structure

We downloaded data samples for crime from the United Nations official webpage, immigration data from Eurostat and use the World Bank Group API to collect population data.

The following tables show how this data is structured to make the data processing easier.

tps00176_linear_2.csv (Eurostat)	universal fields	”geo”	Geopolitical entity	Time period	Flags
value	metadata of the dataset	ISO of the country	Country name	year	flags about the security and privacy of the data

Table 1: General structure of one of the used files

6.2 Data processing

This sources allow our intended use of their sources as defined in their license texts. These licenses are added in our GitHub repository.

Our data pipeline adheres to the Extract, Transform, Load (ETL) framework.

The Extract phase involves two primary methods: manual acquisition of structured .xlsx and .csv files from institutional portals, and programmatic data ingestion from API-enabled sources using Python.

Following extraction, the raw data is loaded into our Python environment for the Transform stage. During this critical phase, our script is executed to cleanse, normalize, and harmonize the data, ensuring all records conform to our predefined master schema.

Finally, in the Load stage, the resulting clean and transformed data is persistently stored in our database, making it readily available for subsequent analysis and visualization.

For the processing we used the python pandas library, which provides intuitive data processing with Dataframe objects.

6.3 Storage

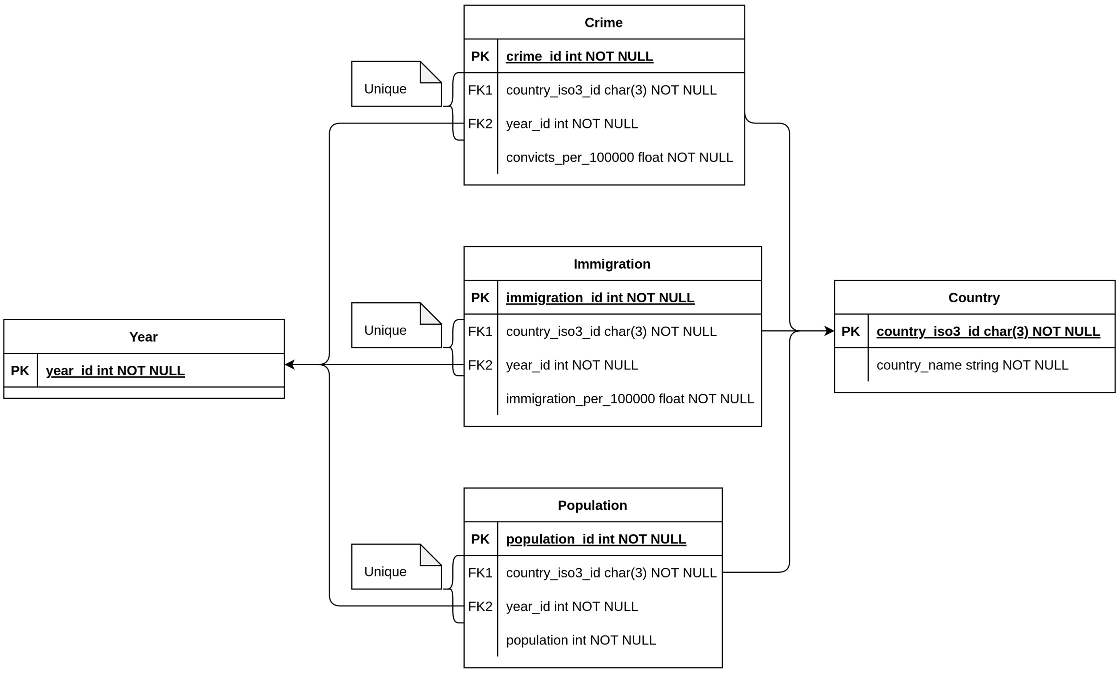
Crime
PK
crime_id int NOT NULL
FK1
country_iso3_id char(3) NOT NULL
FK2
year_id int NOT NULL
convicts_per_100000 float NOT NULL
Unique
Immigration
PK
immigration_id int NOT NULL
FK1
country_iso3_id char(3) NOT NULL
FK2
year_id int NOT NULL
immigration_per_100000 float NOT NULL
Unique
Population
PK
population_id int NOT NULL
FK1
country_iso3_id char(3) NOT NULL
FK2
year_id int NOT NULL
population int NOT NULL
Unique
Year
PK
year_id int NOT NULL
Country
PK
country_iso3_id char(3) NOT NULL
country_name string NOT NULL

Figure 4: Proof of concept ER diagram

Figure 4 shows the real schema of our database. It is based on the previously defined schema draft in the storage chapter. The three important tables of crime, immigration and population statistics are all saved in this database, relating to a specific year and country. We decided to omit information about which types of persons commited crimes, to keep our proof of concept's scope small.

Our PostgreSQL runs in a Docker container, which we can easily backup to save copies of our database.

7 Conclusion

As we do this project, we have come to the realization of the importance that structured data has. It's become clear that having a data source that uses a structured approach can be the difference between a 300 line project and a multiple year endeavor of analysis, crosschecking and statistic inference.

We have also come to the realization that free, open and accessible data is a wonder of humanity we should not take for granted.

This project has served as an exercise to revisit our previous knowledge of databases, while also learning how to work with different systems like APIs. We have had to understand the inner functioning of different organizations to comply with their licensing and know whether a source was reliable, not only in the sense of whether or not we can trust it to have data for a specific field we might need, but also in the sense of whether or not the data offered is true and meaningful.