PLANNING ANALYTIC SOLUTIONS

US Crime Dataset (1980 – 2014)

By Rahul Patel

GitHub
LinkedIn

1. Introduction

Crime remains a major concern in society, impacting public safety, trust, and policy development. By analyzing crime data over time, we can better understand where and why certain types of crimes occur and how effectively they are being addressed.

This project uses a comprehensive dataset covering U.S. crime records from 1980 to 2014, offering over 630,000 entries with information on locations, demographics, crime types, and case outcomes. The goal is to uncover key trends, evaluate crime-solving effectiveness, and explore the social dynamics between victims and perpetrators.

Through data analysis and visualization, this study aims to provide meaningful insights into historical crime patterns and highlight the role of data in informing safer and smarter communities.

2. About the Dataset

2.1 Dataset Description

The dataset used for this analysis consists of **638,454 records** with **24 columns**, capturing a wide variety of attributes related to crimes committed across the U.S. over a 34-year period.

Key Fields Include:

- **Record Identifiers**: RecordID, Year, City, State
- **Crime Details**: Crime Type, Weapon Used, Case Status
- **Demographics**: Age, Sex, Race, Ethnicity of victims and perpetrators
- **Social Context**: Relationship between victim and perpetrator
- **Institutional Data**: Agency involved in case investigation

This rich dataset provides a strong foundation for regional, temporal, and demographic crime analysis.

2.2 Why we selected this dataset

- Crime is a critical societal issue with wide-ranging effects on public safety, policy, and community well-being. Analyzing crime trends helps uncover systemic challenges and regional disparities.
- The dataset spans **34 years (1980–2014)**, enabling long-term trend analysis across different political, economic, and social eras in the U.S.
- It includes demographic, geographic, and contextual information (e.g., victim-perpetrator relationships, weapons used, and case status), making it suitable for multi-dimensional analysis.
- With over **630,000 records**, the dataset is ideal for building dashboards, visualizations, and performance metrics that illustrate real-world data modelling and analysis techniques.

3. Objectives of the Analysis

The main objective of this project is to leverage historical crime data to uncover patterns and generate actionable insights that can inform public safety decisions. By analyzing the dataset, the project aims to achieve the following:

1. Analyze National and Regional Crime Trends

- Examine how crime rates have evolved across the United States from 1980 to 2014.
- Identify fluctuations in violent and non-violent crimes over time.
- Understand how crime patterns vary between states and regions.

2. Identify States with Consistently High or Low Crime Rates

- Rank states based on overall crime volume and crime type.
- Determine which regions have persistent crime challenges and which have improved over time.
- Highlight areas where policy interventions may be needed most.

3. Evaluate the Effectiveness of Law Enforcement

- Measure crime-solving rates to assess the performance of investigative processes.
- Compare case closure success between regions and over different time periods.
- Identify trends in unsolved or incomplete cases to reveal systemic gaps.

4. Key Research Questions

To guide the analysis and ensure a focused and meaningful exploration of the data, the following key research questions were developed. These questions address the most critical aspects of crime trends, geographic disparities, and social dynamics:

1. Which U.S. states consistently report the highest and lowest crime rates?

This question helps identify long-term regional patterns in crime distribution. Understanding which states have persistent crime challenges and which have successfully reduced crime can reveal the influence of local policies, policing practices, socio-economic conditions, and community engagement. It also helps highlight where targeted interventions may be most effective.

2. How have violent crime rates evolved over time from 1980 to 2014?

Crime is not static, it shifts with changes in society, economy, technology, and governance. By tracking violent crime trends over three decades, this question aims to uncover periods of escalation or decline and correlate those shifts with possible causes, such as law enforcement reforms, socio-political events, or economic cycles. Identifying these trends is essential for long-term planning and preventive strategies.

3. What patterns exist in victim-perpetrator relationships, and how do they vary by region or crime type?

This question explores the social dimension of crime. By analyzing the relationships between victims and offenders, whether family members, acquaintances, or strangers, we can gain insights into the nature of violent incidents. Understanding these dynamics, especially in the context of regional and demographic variations, can inform social services, public awareness campaigns, and support for at-risk populations.

5. Defining KPI's

Three primary KPIs were selected to evaluate and compare different aspects of crime data:

1. Incomplete Case Rate

Proportion of records with missing values, particularly in fields such as perpetrator identity, race, or case outcome. High rates may indicate deficiencies in reporting or investigative procedures.

2. Crime Solving Rate

Percentage of crimes successfully resolved, measuring law enforcement effectiveness. This rate highlights both regional disparities and the impact of policing strategies.

3. Crime Growth Rate

Annual rate of increase or decrease in reported crimes, showing national and state-level trends over time.

These KPIs were implemented in the dashboard and served as filters in the visual exploration of crime patterns.

6. Methodology & Data Processing

To ensure the accuracy, reliability, and interpretability of the dataset, a structured methodology was followed. This involved multiple stages of data preparation, transformation, and visualization using industry-standard tools. Below is a breakdown of the key processes:

1. Data Cleaning

The initial dataset contained over 630,000 records, with numerous inconsistencies and missing values across several key fields such as race, age, case status, and perpetrator information. The data cleaning process involved:

- **Removing duplicates** to ensure each crime record represented a unique event.
- **Standardizing missing values** by replacing inconsistent placeholders (e.g., “unknown”, “?”, “.”) with nulls (NaN) for consistency.
- **Filtering out invalid entries**, such as negative ages or blank state fields.
- **Parsing categorical fields**, such as gender and ethnicity, to a uniform format.

This step was essential for maintaining data integrity and avoiding skewed results during analysis.

2. Data Normalization

Regional analysis required consistency in location data, particularly in the **City** and **State** columns. Normalization included:

- Converting all text to lowercase to prevent case-sensitive mismatches.
- Standardizing state abbreviations and full names (e.g., "CA" to "California").
- Resolving city name variants (e.g., “St. Louis” vs. “Saint Louis”) using fuzzy matching techniques.
- Creating a **region code mapping** to allow for regional aggregation (e.g., Northeast, Midwest, South, West).

This ensured accurate grouping for comparative analysis between states and regions.

3. Missing Data Handling and Imputation

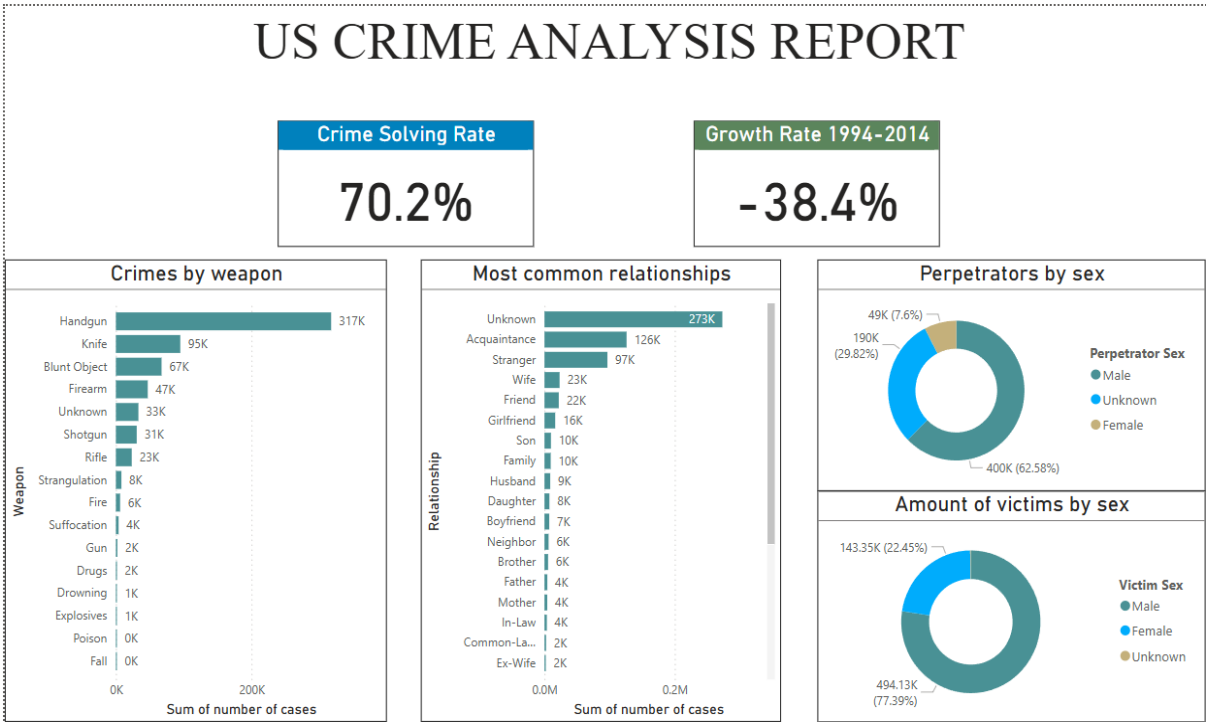
A significant portion of the dataset had missing values, especially in demographic attributes of victims and perpetrators. While not all data could be recovered, imputations were made using logical and statistical methods:

- **Mode or median imputation** for categorical variables with high frequency (e.g., assigning "Male" if 80% of entries were male).

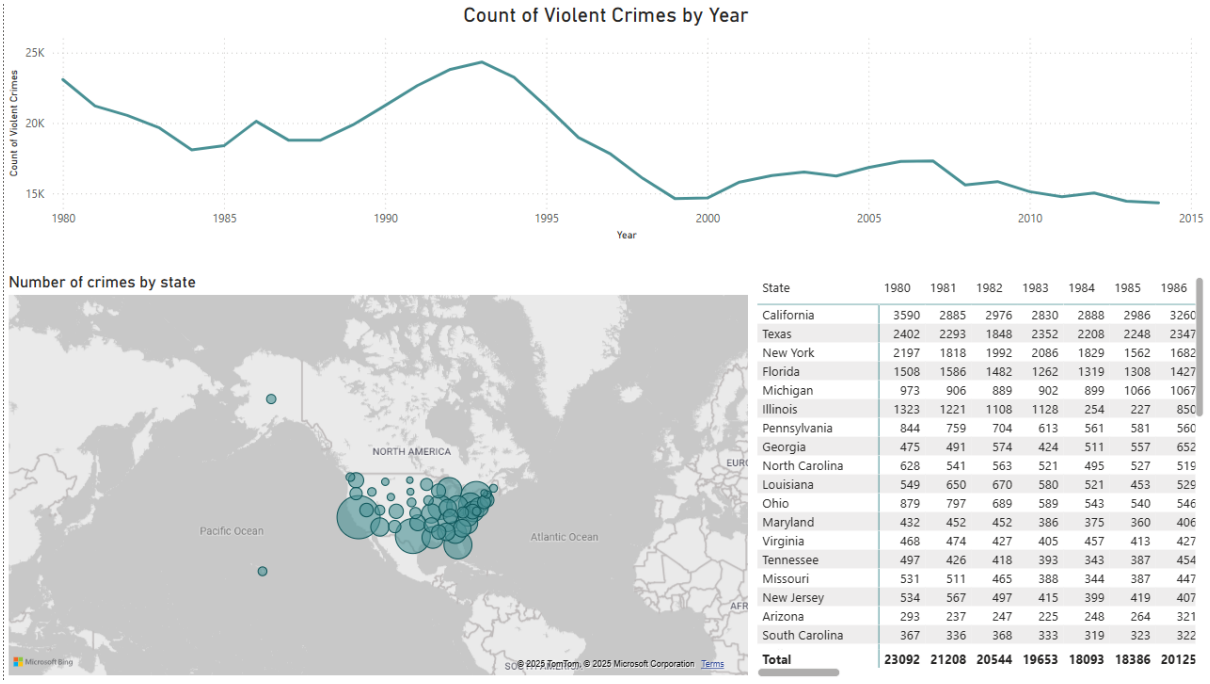
- **Regression imputation** where linear relationships were observed (e.g., inferring perpetrator age from victim age and crime type).
- **Exclusion of outliers** and entries with excessive null values to preserve analytical quality.

This step balanced completeness with caution to avoid introducing bias.

7. Executive Dashboard



8. Reporting Dashboard



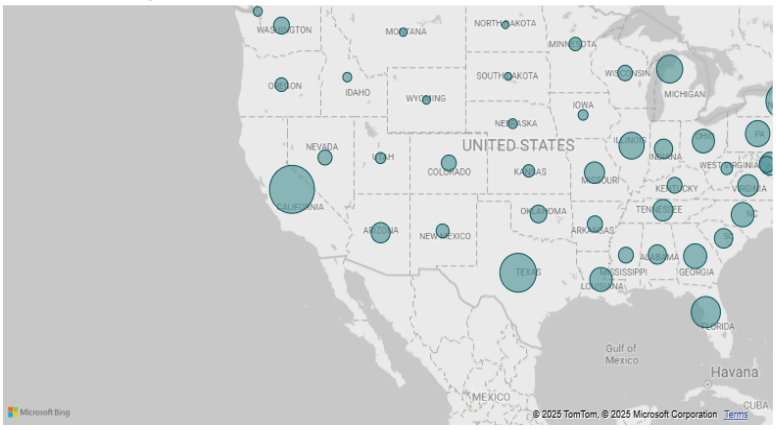
State	1980	1981	1982	1983	1984	1985	1986
California	3590	2885	2976	2830	2888	2986	3260
Texas	2402	2293	1848	2352	2208	2248	2347
New York	2197	1818	1992	2086	1829	1562	1682
Florida	1508	1586	1482	1262	1319	1308	1427
Michigan	973	906	889	902	899	1066	1067
Illinois	1323	1221	1108	1128	254	227	850
Pennsylvania	844	759	704	613	561	581	560
Georgia	475	491	574	424	511	557	652
North Carolina	628	541	563	521	495	527	519
Louisiana	549	650	670	580	521	453	529
Ohio	879	797	689	589	543	540	546
Maryland	432	452	452	386	375	360	406
Virginia	468	474	427	405	457	413	427
Tennessee	497	426	418	393	343	387	454
Missouri	531	511	465	388	344	387	447
New Jersey	534	567	497	415	399	419	407
Arizona	293	237	247	225	248	264	321
South Carolina	367	336	368	333	319	323	322
Total	**23092**	**21208**	**20544**	**19653**	**18093**	**18386**	**20125**

9. Geographic Distribution of Crime

A map-based analysis revealed stark contrasts between states:

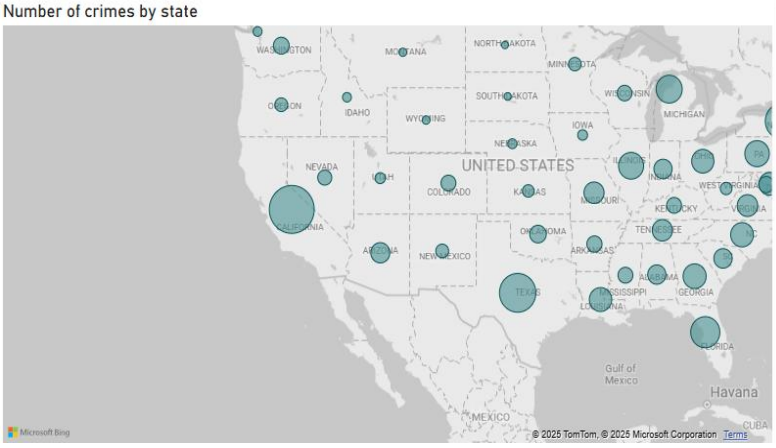
- **High-Crime States**: California, Texas, Florida, and New York, driven by urban populations and socio-economic disparity.
- **Low-Crime States**: Vermont, North Dakota, and South Dakota, likely due to smaller populations, rural settings, and stronger community structures.

Number of crimes by state



State	1980	1981	1982	1983	1984	1985
California	3590	2885	2976	2830	2888	2986
Texas	2402	2293	1848	2352	2208	2248
New York	2197	1818	1992	2086	1829	1562
Florida	1508	1586	1482	1262	1319	1308
Michigan	973	906	889	902	899	1066
Illinois	1323	1221	1108	1128	254	227
Pennsylvania	844	759	704	613	561	581
Georgia	475	491	574	424	511	557
North Carolina	628	541	563	521	495	527
Louisiana	549	650	670	580	521	453
Ohio	879	797	689	589	543	540
Maryland	432	452	452	386	375	360
Virginia	468	474	427	405	457	413
Tennessee	497	426	418	393	343	387
Missouri	531	511	465	388	344	387
New Jersey	534	567	497	415	399	419
Arizona	293	237	247	225	248	264
South Carolina	367	336	368	333	319	323
Total	**23092**	**21208**	**20544**	**19653**	**18093**	**18386**

Number of crimes by state



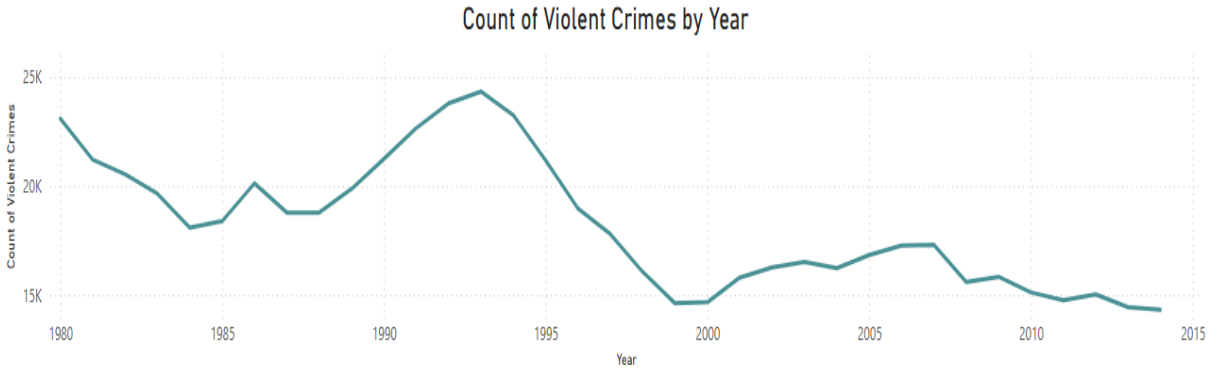
State	1980	1981	1982	1983	1984	1985
North Dakota	9	15	6	15	13	7
Vermont	9	9	5	6	14	17
South Dakota	5	10	16	11	10	9
Montana	29	23	1	23	32	30
Wyoming	34	26	34	30	20	25
New Hampshire	24	28	21	19	12	18
Maine	36	36	25	29	21	29
Idaho	30	45	25	36	37	24
Delaware	44	41	36	27	24	32
Rhodes Island	43	42	36	32	32	36
Nebraska	58	57	35	44	54	46
Hawaii	84	46	33	57	34	43
Alaska	47	69	75	74	52	50
Iowa	63	79	65	70	56	57
Utah	53	54	55	56	49	53
West Virginia	134	122	107	97	88	75
Kansas	163	165	144	146	96	128
Minnesota	112	87	99	68	76	90
Total	**23092**	**21208**	**20544**	**19653**	**18093**	**18386**

10. Trends in Violent Crime

A time-series analysis of violent crime yielded the following insights:

- **1980s**: Violent crime surged, potentially due to economic stress and the crack cocaine epidemic.
- **1990s**: A turning point with steady declines linked to expanded law enforcement, community policing, and technological aids like 911 systems and DNA evidence.
- **2000s to 2014**: Continued decrease in violent crime but with some fluctuations, indicating regional variance and emerging challenges such as cybercrime.

The trendline analysis reinforced the narrative that sustained crime reduction is possible through policy, investment, and data-driven interventions.



11. Demographic & Social Analysis

Perpetrators

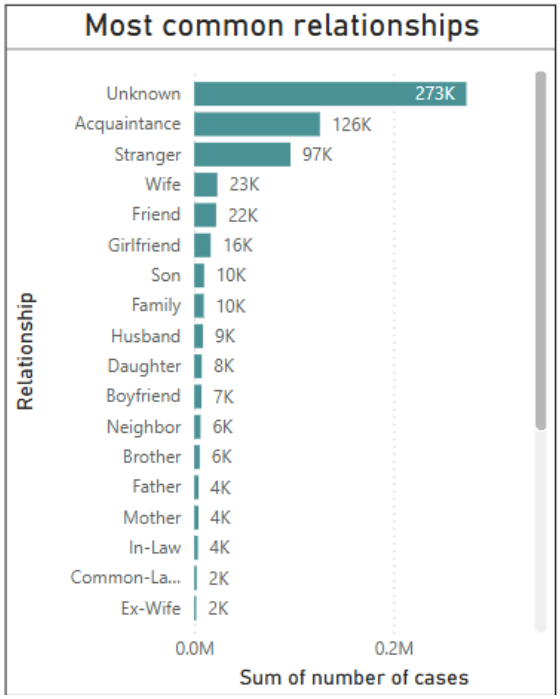
- Predominantly **male**, with a concentration in the **18–35** age group.
- Racial patterns varied by region, but data quality in some demographic fields limited comprehensive analysis.

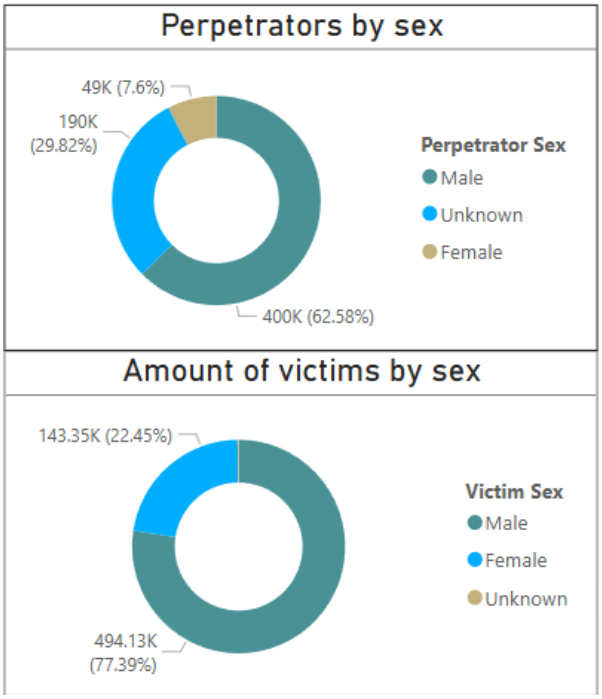
Victims

- Similar age and gender patterns to perpetrators.
- Female victims were disproportionately represented in crimes involving known perpetrators (e.g., domestic abuse).

Relationships

- Over 40% of crimes involved a known relationship between the victim and perpetrator.
- Family violence, intimate partner abuse, and acquaintance-based crime were prevalent in non-urban areas, suggesting under-reported social dynamics.





12. Observation & Key Takeaways

Through detailed analysis of the dataset, several important patterns and insights emerged. These observations not only reflect the trends captured in the data but also reveal systemic factors that influence crime reporting, resolution, and prevention in the United States.

1. Crime Is Not Evenly Distributed

Crime patterns vary significantly across states and over time. Larger, urbanized states such as California, Texas, and New York report higher crime volumes, while rural or less populated states like Vermont and North Dakota show consistently low figures. This uneven distribution is influenced by several factors:

- **Population density**: More densely populated areas tend to have higher crime rates due to increased interaction and anonymity.
- **Urban vs. rural environments**: Urban centres typically experience more property and violent crimes, while rural areas may report fewer incidents or different types of crime.
- **Socioeconomic disparities**: Regions with high poverty rates, lower educational attainment, and limited access to healthcare or employment are more vulnerable to crime.

These geographic variations underscore the need for **localized policy approaches** instead of universal strategies.

2. Relationships Matter

One of the most revealing insights from the data is the significant proportion of crimes that occur between individuals who know each other. Rather than being random acts of violence, many crimes involve:

- Family members
- Friends or acquaintances
- Coworkers or neighbours

3. Law Enforcement Efficiency Varies Widely

The crime-solving rate—i.e., the percentage of cases marked as resolved or closed—differs significantly between states and across crime types. Several observations support this:

- **High-crime urban areas** often show lower resolution rates, potentially due to limited resources, case backlogs, or lack of witness cooperation.

- **Well-funded departments** with access to technology (e.g., forensic labs, data systems) tend to perform better in solving crimes.
- **Disparities in case management** and investigative procedures across jurisdictions suggest a lack of standardization.

13. Limitations

1. Data Quality and Completeness

One of the most significant challenges was the inconsistency and incompleteness of many records. Several fields, especially those related to **demographics (age, race, sex)** and **case status (solved/unsolved)**, contained missing or ambiguous values.

- Certain crimes, particularly older ones or those occurring in rural jurisdictions, lacked detailed entries on the perpetrator or victim.
- In some instances, key fields such as weapon used or relationship were missing entirely, limiting the ability to perform deeper relational analysis.

This limitation reduced the analytical precision in demographic studies and introduced uncertainty into certain KPI calculations.

2. Reporting Bias and Inconsistencies

The dataset likely suffers from **systematic underreporting** and **inconsistent data collection practices** across states and time periods. This is a well-documented issue in criminal justice data and affects interpretation in several ways:

- Not all crimes are reported to authorities especially domestic violence, sexual assault, and other sensitive crimes leading to data gaps.
- Law enforcement agencies differ in classification methods, which means that what is reported as “assault” in one state might be categorized differently in another.
- Some jurisdictions may prioritize different types of data collection, causing variation in what gets recorded in the database.

As a result, while the dataset offers a large and valuable sample of crimes, it does not represent the totality of criminal activity in the U.S., nor does it provide uniform coverage across all regions.

Final Reflection on Limitations

Despite these challenges, the dataset remains a **valuable lens into historical crime patterns**, offering insight into decades of criminal activity, justice system effectiveness, and public safety evolution in the U.S. While imperfect, the data still enables meaningful analysis—particularly when its limitations are acknowledged and addressed through careful data handling and contextual interpretation.

14. Future Additions

While this project provides meaningful insights into U.S. crime data from 1980 to 2014, there is substantial potential for further development and enhancement. Future work can expand the scope, analytical depth, and real-world utility of this project by incorporating more advanced methods, recent data, and interdisciplinary perspectives.

- Incorporate **post-2014 data** from FBI UCR or BJS.
- Apply **machine learning** for crime type prediction or perpetrator profiling.
- Add **economic and educational indicators** for a socio-economic overlay.
- Integrate **real-time feeds** to build a dynamic, auto-updating dashboard.

These enhancements would create a living tool for researchers, law enforcement, and policy professionals.

15. Conclusion

Analyzing U.S. crime data from 1980 to 2014 shows how data can help us better understand social issues like crime and safety. This project uses clear visualizations, thoughtful analysis, and real-world context to highlight patterns and trends that matter.

It also shows the importance of open data and how data science can support smarter decisions, better policies, and stronger communities.