Assignment 1

Student Name: Guadalupe María Armenta Mendoza

Student No: R00259577

For the module DATA9005 as part of the

Master of Science in Data Science and Analytics, Department of Mathematics, 2024/25

Declaration of Authorship

I, Guadalupe Armenta, declare that the work submitted is my own.

- I declare that I have acknowledged all main sources of help
- I declare that I have not used ChatGPT, genearative AI software, or other similar software in this assignment.
- I declare that I have not used a language translation software in the completion of this examination.
- I declare that I have not obtained unfair assistance via use of a third party.
- I acknowledge that the Academic Department reserves the right to request me to present for oral examination as part of the assessment regime for this module.
- I confirm that I have read and understood the policy and procedures concerning academic honesty, plagiarism, and infringements.
- I understand that where breaches of this declaration are detected, these will be reviewed under MTU (Cork) policy and procedures concerning academic honesty, plagiarism, and infringements, as well as any other University regulations and policies which may apply to the case. I also understand that any breach of academic honesty is a serious issue and may incur penalties.
- Examination/assessment material may, at the discretion of the internal examiner, be submitted to the University's plagiarism detection solution.
- Where I have consulted the published work of others, this is always clearly attributed using appropriate citations and references.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this work is entirely my own work.
- I acknowledge the use of Grammarly.com in helping me to review my writing and polishing at the final stage of preparing my assessment. I used the following prompt: "suggest ways to improve clarity and concision. Provide general advice with examples.

Signed:
Date: 09/03/2025

Tabla de contenido

1. Introduction

This project aims to analyze crime trends in Mexico from 2015 to 2024 using data visualization techniques. By using open crime datasets and integrating them with additional sources such as population statistics and socioeconomic indicators, this study provides a more comprehensive perspective on crime distribution and its potential correlations with social factors.

This assignment uses the open dataset "Cifras de Incidencia Delictiva Estatal" which covers crime data from January 2015 to December 2024. The dataset is sourced from the official governmental platform "Datos Abiertos de Incidencia Delictiva" (Gobierno de México, 2025). It provides categorized records of crime incidents, classified by crime type, subtype, and method of commission. It also includes detailed information on the time (month), frequency of crimes, and state-level distribution. Since crime counts are recorded monthly, this data allows for a precise analysis of trends and patterns over time.

However, it is important to note that measuring Mexico's actual crime rate is challenging due to substantial underreporting in official records. According to government surveys like the Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE) only about 10% of crimes are reported to authorities (INEGI, 2020). “Many Mexicans are reluctant to report crimes because of a lack of trust in the police and the justice system, along with concerns about possible retaliation from criminal organizations” (Valle, D. 2023). This limitation must be considered when interpreting the results of this study.

2. Questions

a) Where did you get the idea/code for this concept/implementation?

I obtained the idea for the map visualization from the OCVED (Organized Criminal Violence Event Data) website, which shows the territorial presence of various cartel organizations in Mexico between 2000 and 2019 (Osorio, 2020).

I also drew inspiration from the Elcrimen website, which focuses on five key crimes—kidnappings, homicides, extortions, femicides, and car theft (Valle D., 2015). Although the author provides a very useful R code for the map and time series, I decided to use Python for my project because it was easier for me to code.

I used a GitHub repository for the Mexican map provided by Valle D., (2023) and completed courses on Visualizing Geospatial Data in Python and Intermediate Data Visualization with Seaborn on DataCamp. In addition, I watched the Plotly Tutorial 2023 by Derek Banas to learn how to create interactive time series and maps. This was essential for me in producing most of my plots.

b) Has anyone else published visualisations (on the internet etc.) for this dataset?

Yes, other visualizations have been published using this dataset. For example, Valle D., (2015) regularly updates visualizations on Elcrimen website that focus on five types of crimes and their development across different Mexican states. However, given that the dataset contains information on forty crime types and Mexico is divided into thirty-two federal entities (thirty-one states and Mexico City), creating comprehensive plots for every crime and every state would be very complex. My approach, therefore, focuses on a broader analysis that balances detail with clarity.

c) What is novel about your approach in this assignment?

I chose to perform a deeper analysis of the primary crime dataset by merging it with three additional data sources:

Poverty Indicators (2008–2018): Provided by CONEVAL, this dataset includes social factors such as healthcare, social security, educational lag, poverty, and extreme poverty. I used these indicators to explore if deficits in social services correlate with crime rates in 2018. A significant challenge was that the poverty data only matched with the crime data for the year 2018. Consequently, I was able to analyze the relationship between crime and these social factors only for that specific year. I addressed this by creating a correlation matrix to evaluate how each social factor is associated with the total number of crimes.

Population Data (2015–2024): Sourced from The World Bank Group, this data enabled me to standardize crime numbers per 100,000 inhabitants, thus offering a clearer comparison across years. I visualized this data using a bar plot that illustrates the annual crime rate per 100,000 inhabitants.

Population and Housing Census (2020): From INEGI, this dataset allowed for a detailed state-level analysis of crime rates and made it possible to compare the top five most dangerous states with the bottom five of 2020. From this dataset, I generated multiple visualizations, including a Crime Rate Map by State in Mexico (2020), a bar chart comparing the Top 5 Most and Least Dangerous States (2020), and a vertical bar chart of State Crime Rates (2020).

This multidimensional approach provides a more informative view of crime by linking it with social and economic factors, which I believe is both innovative and useful for further research.

d) Detail at least one complex concept/implementation in your work?

One of the most challenging aspects was merging the main crime dataset with the poverty indicators dataset (2008–2018, national and state-level). This second dataset presented several difficulties: the state names differed (including the use of accents), the variable types did not match, and the structure was reversed, years were columns in the poverty dataset, while in the crime dataset they were rows.

To unify these datasets, I first had to match the state names and standardize the variable types as they were assigned in the original files. I also needed to remove commas and convert the relevant fields to numeric types (ensuring that columns were treated as strings first), then rename the columns and translate them into English, as they were originally in Spanish. Honestly, if I had known that this unification would not lead to interesting or useful plots, I might have reconsidered using this second dataset.

Another complex challenge involved the main crime dataset itself, where all the information was in Spanish. Translating variables—especially transforming the "Crime type" field into English, which involved 41 different crime categories—proved to be quite time-consuming.

Additionally, creating the interactive map (Plot 6) was challenging. I had to consult Plotly's documentation to understand how to properly integrate a JSON file into this type of chart. At one point, I used ChatGPT to diagnose the error I was encountering and learned that to use a JSON file with Plotly, it needed to be converted to GeoJSON format.

Finally, when developing the map for crime rates by state (Plot 8), I found the GitHub resource by Valle, D (2023) and his website extremely useful. His work provided clear guidance on how to integrate the data to create the map, which was instrumental in overcoming some of the technical hurdles.

Overall, the challenges of data cleaning, transformation, and integrating multiple data formats required the most time and involved consulting different websites

e) Discuss at least three theoretical/design consideration/choice you made.

When designing the visualizations for this project, I applied several theoretical and design principles from well-known data visualization experts discussed in class, including Nathan Yau and Edward Tufte. Below are three design considerations that guided my approach:

Tufte's First Principles:

A core principle in Edward Tufte's five principles of design is the importance of showing comparisons (Tufte 2019). To adhere to this, I implemented a horizontal bar plot (Plot 7), which directly compares the top 10 most dangerous states in 2020 versus 2024. This allows for a clear visual representation of changes in crime rankings over time, making it easier to infer how crime has evolved across different states. By providing side by side comparisons, the audience can quickly assess regional crime trends and how they shift over time.

Another key design choice was ensuring that my visualizations helped answer important questions, as emphasized by Tufte (2019). For instance, Plot 8 (Top 5 Most and Least Dangerous States in 2020) helps address the question of which states had the highest and lowest crime rates. Similarly, Plot 2 (Monthly Crime Trends) explores seasonal patterns in crime rates, helping us infer whether certain times of the year are more prone to criminal activity. These choices align with the idea that data visualizations should display information and provide insights that help the audience understand underlying trends and causality.

Using Maps and Color Theory – Nathan Yau's Principles:
Visualization That Means Something by Nathan Yau (2013), highlights the importance of map and color choices in effectively conveying data. Inspired by his work, I incorporated a crime rate map (Plot 6) to visually represent crime distribution across Mexican states. I also paid attention to color schemes, ensuring that higher crime rates were represented with more intense colors to enhance readability and interpretation.

f) ***What aspect of data analysis did you investigate (e.g. pattern recognition, distribution comparison, statistics etc.)?***

During this project, I explored multiple aspects of data analysis to gain a comprehensive understanding of crime trends in Mexico. The main analytical approaches are the following;

- The time series plots (Plots 1 and 2) helped identify trends in crime rates over time (year and monthly), allowing us to infer periods of increase or decline, including the drop in crime during 2020, likely due to the COVID-19 pandemic.
- Some visualizations, such as Plot 7 (Top 10 Most Dangerous States in 2020 vs. 2024), were designed to compare different periods or regions. The same comparison idea is followed by Plots 3, 4 & 5. This approach helped highlight which states or crime type improved or worsened in terms of safety across 2015 to 2024.
- The correlation matrix (Plot 9) examined the relationship between crime rates and socioeconomic factors such as education, healthcare access, and poverty. This helped infer potential underlying causes of crime.
- By integrating a geographical component in Plot 6 (Crime Rate Map) and Plot 9, I visualized how crime varies across different regions of Mexico, revealing spatial patterns and clusters of high-crime areas.

g) Given more time, how can your work be improved?

With more time, several improvements could be made:

An interesting observation in Plot 8 is that Colima appears as one of the states with the highest crime rate per 100,000 inhabitants in 2020, even though Mexico City recorded the highest absolute number of crimes that year, which is unusual. A brief online investigation revealed that Colima experienced significant clashes between police and narco groups (Osorio, D, 2020), which likely impacted on its crime rate. With more time, I would explore this anomaly further by analyzing additional newspapers and websites to better understand the underlying factors.

Firstly, investigating some of the unusual findings I encountered, for instance, in Plot 5, while robbery is the most frequently committed crime, crimes such as child trafficking, abduction, and feminicide have low occurrence rates despite extensive media coverage in Mexico. Further research is needed to understand these discrepancies.

Secondly, the correlation matrix (Plot 9) indicates that a lack of access to healthcare is strongly linked with higher crime rates. More investigation is required to determine whether these factors are causal or merely associated.

The third improvement is related to the process of manually creating regression plots to examine the relationship between social factors and crime for each of the 32 entities (31 states plus Mexico City) is time-consuming. Automating this process would enable a more in-depth and efficient analysis. Furthermore, I would like to explore advanced visualization techniques, such as interactive 3D plots, to present this data more engagingly rather than using simple regression plots.

Additional time would allow for refining the data merging process. Incorporating more granular datasets, such as local law enforcement reports or community surveys, would help address issues like underreporting and provide a more accurate picture of public safety.

Moreover, I believe that my data visualization plots lack well-structured development and more engaging color schemes. The plots tend to be somewhat repetitive, which is not ideal for engaging the audience and presenting the data in an informative manner. These improvements would not only deepen the analysis but also improve the overall presentation and clarity of the findings.

3. Conclusion

This project aimed to analyze crime trends in Mexico from 2015 to 2024 using data visualization techniques. By merging the main crime dataset with population and socioeconomic indicators, I was able to provide a more comprehensive perspective on how crime interacts with broader social factors.

The visualizations revealed key trends, such as the decline in crime during 2020, likely linked to the pandemic, and the persistence of high crime rates in specific states like Mexico City and the State of Mexico. The correlation matrix also suggested a strong relationship between crime and deficiencies in education and healthcare, reinforcing the idea that improving social conditions may play a role in reducing crime.

Ultimately, this project highlights the power of data visualization in uncovering crime patterns and providing insights that could be valuable for public policy and law enforcement strategies.

However, there is still room for improvement in terms of refining the color schemes, increasing the diversity of visualization types, and automating certain analyses for efficiency.

Future work could involve refining the visualizations further, integrating additional datasets, and exploring predictive modeling techniques to anticipate crime trends based on historical data.

Word Count: 2,162
Harvard Style

4. Reference List

Banas, D. (2020). *Plotly Tutorial 2023*. [online] www.youtube.com. Available at: https://www.youtube.com/watch?v=GGL6U0k8WYA [Accessed 16 Feb. 2025].

CONEVAL ed., (2019). *Indicadores de pobreza, 2008-2018 (nacional y estatal) - datos.gob.mx/busca*. [online] Datos.gob.mx. Available at: https://datos.gob.mx/busca/dataset/indicadores-de-pobreza-2008-2018-nacional-y-estatal [Accessed 15 Feb. 2025].

INEGI (2021). *Censo Población y Vivienda 2020*. [online] www.inegi.org.mx. Available at: https://www.inegi.org.mx/programas/ccpv/2020/#Tabulados [Accessed 20 Feb. 2025].

INEGI ed., (2024). *Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE) 2024*. [online] Inegi.org.mx. Available at: https://www.inegi.org.mx/programas/envipe/2024/ [Accessed 3 Mar. 2025].

Kabacoff, R. (2020). *Data Visualization with R*. [online] *rkabacoff.github.io*. CRC Press. Available at: https://rkabacoff.github.io/datavis/ [Accessed 9 2025].

Osorio, D. (2020a). ¿Qué pasa con el homicidio en Colima? *Nexos.com.mx*. [online] doi:https://doi.org/10883332/article-600x250-ros.

Osorio, J. (2020b). *Organized Criminal Violence Event Data.(OCVED)*. [online] OCVED. Available at: https://www.ocved.mx [Accessed 20 Feb. 2025].

Plotly (2023). *Plotly Python Graphing Library*. [online] plotly.com. Available at: https://plotly.com/python/ [Accessed 20 Feb. 2025].

The Met (2009). *Pen and Parchment - The Beautiful Evidence of Medieval Drawings*. [online] YouTube. Available at: https://www.youtube.com/watch?v=HfXSltlDfDw [Accessed 20 Feb. 2025].

Valle, D. (2015). *Reporte criminal de enero 2025* . [online] https://elcri.men/. Available at: https://elcri.men/ [Accessed Feb. 15AD]. Todos los datos de este sitio web provienen del SESNSP y el INEGI (sólo homicidios).

Valle, D. (2023). *GitHub - diegovalle/mxmaps: An R package for making maps of Mexico*. [online] GitHub. Available at: https://github.com/diegovalle/mxmaps [Accessed 20 Autumn 2025].

Yau, N 2013, Data Points: Visualization That Means Something, John Wiley & Sons, Incorporated, Newark. Available from: ProQuest Ebook Central. [9 March 2025].

World Bank (n.d.). *Population, total | Data*. [online] Worldbank.org. Available at: https://data.worldbank.org/indicator/SP.POP.TOTL?locations=MX [Accessed Feb. 2015].