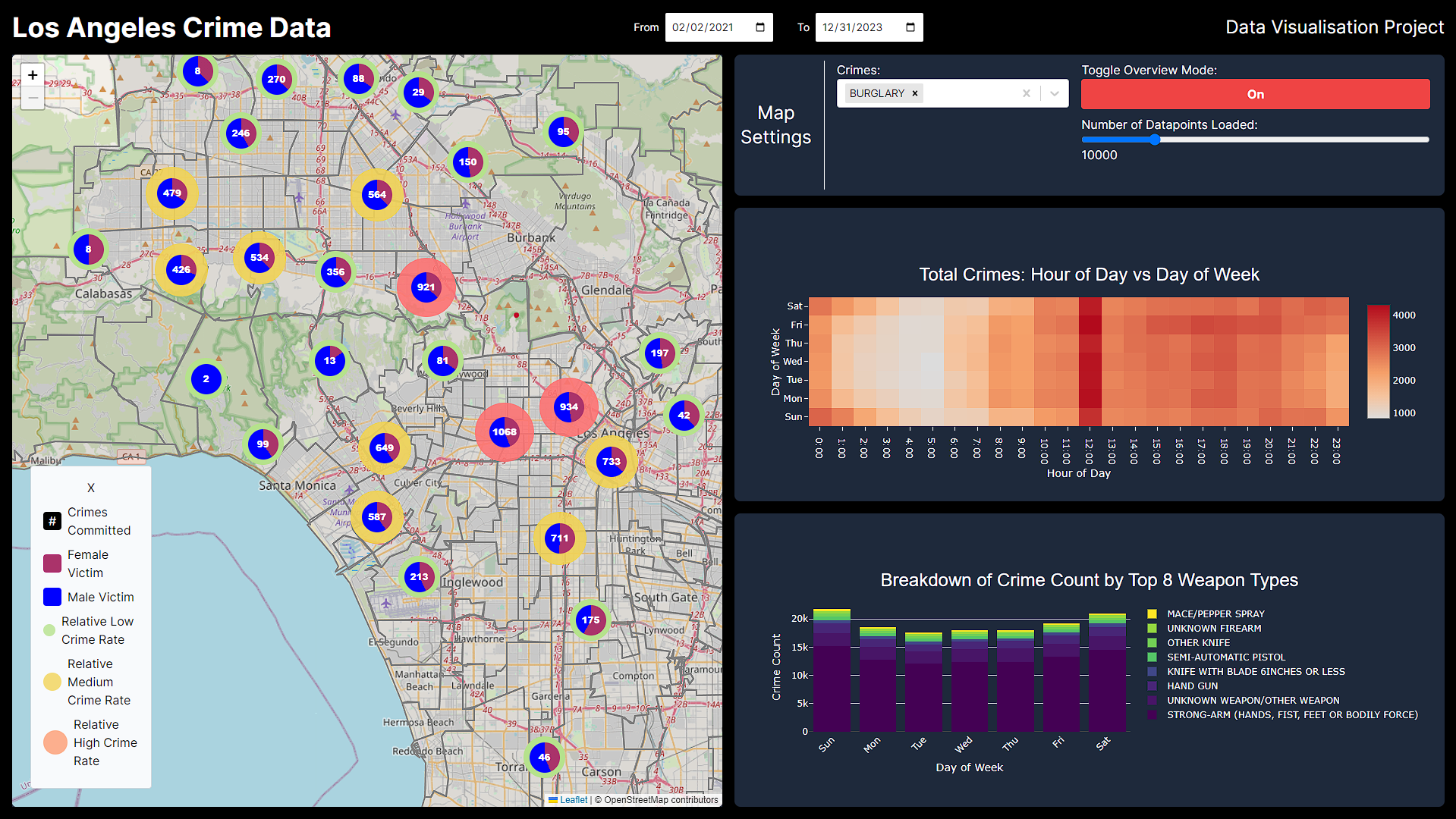
****

**DECLARATION:** I understand that this is an **individual** assessment and that collaboration is not permitted. I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at <http://www.tcd.ie/calendar>. I understand that by returning this declaration with my work, I am agreeing with the above statement.

# Dataset

The dataset visualised in this assignment is the Los Angeles crime data from 2020 to 2023. The data was obtained from Kaggle (1). The dataset is substantial, there are ~700,000 rows of data in the original file. For the purpose of this assignment, the dataset was cleaned to remove rows with null values in critical columns such latitude, longitude, crime committed, etc. After cleaning, the dataset had ~500,000 rows. There was 28 columns of data included in the raw dataset. Some of the columns can be derived from others, such as time between crime committed and date reported, and others are not interesting in terms of visualisation potential. As a result, the number of columns were reduced to 9.

|  |  |  |
| --- | --- | --- |
| Column Used | Data type | SQL type |
| **Date occurred** | Temporal | Date |
| **Time occurred** | Temporal | Time |
| **Crime code** | Ordinal | Int |
| **Crime description** | Nominal | Varchar |
| **Victim age** | Discrete | Int |
| **Victim gender** | Nominal | Boolean |
| **Latitude** | Ratio | Decimal |
| **Longitude** | Ratio | Decimal |
| **Weapon used** | Nominal | Varchar |

The dataset is tabular in nature and each row of the table represents a crime that was reported to the police. From basic analysis, it appears that it is not a comprehensive set of data for the Los Angeles area. Many parts of Los Angeles have no crimes reported, leading me to believe that the dataset is incomplete. However, for the purposes of this assignment, the dataset contains enough information to create meaningful visualisations.

Due to the volume of data present, it would be impossible to perform any meaningful analysis without visualisations. The aim of this particular visualization is to empower the viewer to form their own conclusions. For decision-makers within the Los Angeles Police Department or informed citizens in LA, it could hypothetically facilitate potentially life-saving decisions.

Within the heatmap and the stacked bar chart, an ordinal cyclical attribute of the day of the week over the time period specified in the top bar.

# Tools / Technologies used

Before creating the visualisation, the data needed to be cleaned. As mentioned above, the data in its original state had almost 700,000 rows. The cleaning process involved the use of a Python script which looked for columns that had null or malformed data. After this process, the dataset reduced to roughly 500,000 rows.

The visualisation itself comes in the form of a web-based dashboard. This was created using Next JS (2) as the framework. This allowed for both frontend and backend development to be done within the single repository. The frontend itself is written in React JS (3), and the backend application programming interface endpoints are written in JavaScript. Due to the large volume of data, the browser was not capable of handling the memory usage. As a result, a SQLite (4) database was implemented to make querying faster and easier. A script, written in Python, was created to transfer the data from CSV format to the database. Another script, written in JavaScript to run in Node JS (5), was created to initialise the database and add indexes to speed up querying times. Leaflet JS (6) was utilised for the mapping functionality. All other graphs were created using Plotly JS (7).

# Tasks

Before implementing the visualisation, two imaginary viewer profiles were chosen to determine the tasks that would be required. The first was a decision maker within the Los Angeles Police Department. This individual is hypothetically responsible for personnel planning and allocation of resources within the police department. The second profile is that of a conscious individual who wishes to be better informed regarding the safety of different areas across Los Angeles.

## Map

* Explore the most common areas for crime across the city. The numbers within the markers refer to the total number of crimes committed within that cluster. This will give an indication of relative danger given a specific area. It is also possible to filter by specific crimes if that is a requirement.
* Identify correlations between area and type of crime committed, number of crimes committed, type of weapons used, etc. Within each of the clusters, there are further points. When one gets down to the individual crime level, it is possible to see further details related to the crime listed before.
* Lookup the types of crime in a specific area.
* Compare victims of crimes by gender and the area in which the crime occurred.

## Heat Map

* Compare and contrast volume of crimes against the time of day across each day of the week. This will allow the viewer to discover what days and times of the week have the highest crime incidence rate.
* Summarize crime committed per day by each hour.
* Query specific day and time combinations to find the number of crimes committed (using the tooltip to get specific numbers).

## Stacked Bar Chart

* Explore the top weapons used by the total number of crimes committed with that weapon. This could help inform legislative decision making around what weapons are of particular threat to the people.
* Query weapon usage on particular days.
* Search for the most common weapon type.

# Encoding Channels and Idioms

## Idioms

The visualisation is made up of various sub plots that can be controlled using a centralised dashboard settings panel. The 3 component plots share the same date range that can be controlled from the top bar.

The map of Los Angeles plot is the main component of the visualisation. It takes the common spatial idioms that plots dot density on the map, and clusters them together into groups. The clusters are grouped together based on their proximity to one another, and a weighted centre is selected. The pie chart idiom is employed to convey the proportion between male and female crime victims. The markers themselves are located at the latitudinal and longitudinal coordinates at which the crime occurred. The ring around each marker, that has a colour of green / yellow / red, uses a variant of the heatmap idiom. This is for easy identification of areas that have a high crime rate. Hovering over one of the clusters causes the area in which the cluster represents to be shaded using a coloured polygon. This could be considered a form of area-based idiom as it allows to viewer to easily identify the area in which the cluster represents.

The heatmap of number of crimes compared to time of day and day of week is a standard XY heatmap. In this case, the heatmap idiom suited this task perfectly, allowing the viewer to query the data and spot trends with ease.

The final chart is the stacked bar chart of crimes committed compared across each day of the week and what weapon was used. The stacked bar chart idiom allows for a simple way to show part-to-whole relationships. In this case, we can see the relationship between days of the week, weapon used, and crime count.

## Encoding Channels

The map of Los Angeles uses various encoding channels to convey the data in a clear way. The first idiom used is position. Position of the markers indicate the areas that the crime occurred. Due to the underlying data being latitudes and longitudes, position on the map is the perfect encoding channel. Colour is used in various ways. Firstly, it encodes the male versus female split of crime victims on each marker. Secondly, it used when the viewer zooms into the map to distinguish what crime was committed. Finally, it shows the relative amount of crimes committed using either green, yellow, or red depending on the number of crimes. Size is also employed as an encoding channel. It is used to emphasize the safety level of each area under the marker. The larger the size, the more crime that occurs in that area.

The heatmap uses two main encoding channels. The first is the colour of the square. The darker shade of red indicates that more crimes occurred at that time and day. The second encoding channel is the position of the square on the axis. This position indicates the hour of the day and the day of the week.

The stacked bar chart uses the same encoding channels as the heatmap. Colour is used to what the weapon is. The position of the bars show the day of the week and the number of crimes committed.

# Novelty

The novelty in this visualisation can mostly be seen within the map of Los Angeles. Deviating from the conventional spatial idioms, the visualisation utilises a combination of spatial and pie chart idioms. Traditionally, spatial data is presented through dotted points indicating crime incidence. In this case, it converges with the pie chart idiom to highlight the gender disparities among crime victims. This unconventional use of idioms is the key contributor to the novelty of this visualisation.

The format of the visualisation further adds to its novelty. Formatted in a dashboard style, this approach is unique when considering the dataset chosen and the columns chosen for the visualisation. The use of the dashboard format gives the viewer an overview of crime from various angles, enhancing the overall novelty.

# Critical Analysis

## Strengths

Clarity is a key strength of this visualisation. Clutter is avoided by clearly laying out the dashboard in an easy-to-understand way. Within the map, clustering is used to maintain visual clarity and handle the complexity of the large volume of data. Exploration is also incentivised as the viewer can interact with the map, zooming in on areas and seeing what sub-clusters lie within a given area. The use of tooltips help the viewer query specific information when it is necessary.

Within the heat map and the stacked bar chart, graphical excellence is taken into consideration, only showing the data and avoiding the use of extra style or unneeded features. Chart junk is avoided through the use of transparent backgrounds.

## Weaknesses

Due to limited space and the idioms chosen, may of the original columns in the dataset had to be excluded. Although some of the excluded columns were superfluous, there were some that could have proved to interesting if incorporated into the visualisation.

Animation could have been used to visualise the changes over time. Currently, changes over time are only visible if the user manually changes the dates, which can become cumbersome.

# References

1. Los Angeles Crime Data 2020-2023 [Internet]. Available from: https://www.kaggle.com/datasets/venkatsairo4899/los-angeles-crime-data-2020-2023

2. Next.js by Vercel - The React Framework [Internet]. Available from: https://nextjs.org/

3. React [Internet]. Available from: https://react.dev/

4. SQLite Home Page [Internet]. Available from: https://www.sqlite.org/index.html

5. Node.js [Internet]. Available from: https://nodejs.org/en

6. Leaflet — an open-source JavaScript library for interactive maps [Internet]. Available from: https://leafletjs.com/

7. Plotly [Internet]. Available from: https://plotly.com/javascript/