澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

CISC3015 - Data and Information Visualization (2024-1 001)
Project Report

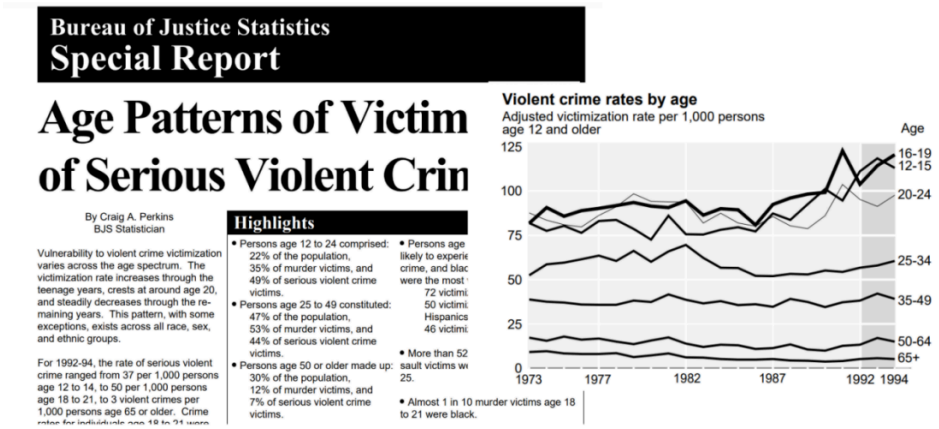
Pair Number: 05
Student ID: DC229805; DC427182
Student name: Zhong Wanyi; Tailine Nonato

Project Background

Our main issue is about what are the factors that explain the number of crimes by category in Los Angeles from 2020 to 2024 as focusing on the specific crime type of criminal homicide. According to a recent study, the research has shown that certain demographic factors like gender and age can influence criminal homicide rates and patterns, hence we decide to assume victim's gender and age as the main factors to affect the number of crimes. In reality, males tend to be more frequently victims of homicide than females across most societies. Age is also a risk factor, with young adults often being disproportionately affected by homicidal violence.

Then, we take one of the previous references as an example. The special report was reported by the Bureau of Justice Statistics, aimed to make a conclusion on age patterns of victims of serious violent crime. Take a look at the final chart mentioned in such a report, it was shown that teenagers had a higher possibility of being victims, while elderly were less likely to be killed by offenders. As a result, the age of the victim should be negatively correlated with the crime rate.

Figure 1 - Related references



Project Goals

Our purpose of doing our pair project is to make some contributions to individuals and society. Individually, our goal is to advance knowledge in fields like data analysis, sociology, psychology related to deviant behaviour. When thinking about the whole society, our goal is to identify high-risk gender and age groups that may need more support services or protection programs to protect their safety in the future.

Project Tasks

(1) Main Task: do a data analysis to conduct the relationship between age, gender and the number of incidents, and find out the high-risk gender and age groups.
(2) Identify the region and the domain category: Los Angeles.
(3) Identify the dataset we require to use: Crime_Data_from_2020_to_Present.csv
(4) Identify the related issue we want to investigate: What are the factors that explain the number of crimes by category in Los Angeles from 2020 to 2024? Does anything change when looking in detail into criminal homicides?
(5) Identify what kinds of visualizations we want to use: count plots and line plots.
(6) Finish our project proposal.
(7) Visual our pair project successfully.
(8) Do exploratory data analysis (EDA).
(9) Make a conclusion.

Project Deliverables

I. Project presentation

The project live presentation was first done for the class and it was a short report of the work done until then. After the live presentation, we worked on improvements to enhance the research question and the methodological approach over it. The now deliverable presentation includes the final research question, visualizations and comparison between the background information and the actual updated data.

II. Dataset

The dataset, which comes from a reliable and direct source, is reported by the Los Angeles Police Department (LAPD). The dataset is intended for public access and use and is covered by different Terms of Use than Data.gov. Although the source is reliable, inaccuracies could happen as the collection of data is still precare. Quoting the LAPD:

> *"This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data."*

To the date we download the dataset, it is shaped into 28 columns and 722,906 observations, offering not sensitive information about the crimes committed as region and crime description and certain characteristics of the victim's.

III. Jupyter Notebook

The Jupyter Notebook format was used to do the needed data analysis in a way that with over 30 chunks of code, visualizations and specific functions were created to optimize and reach the final goals. The programming language used was Python in its 3.12.7 version with the help of the *pandas* and *numpy* libraries.

IV. Visuals

The following 8 visuals were produced by the analysis:

Figure 2 - Percentage of crimes by category and victim's sex

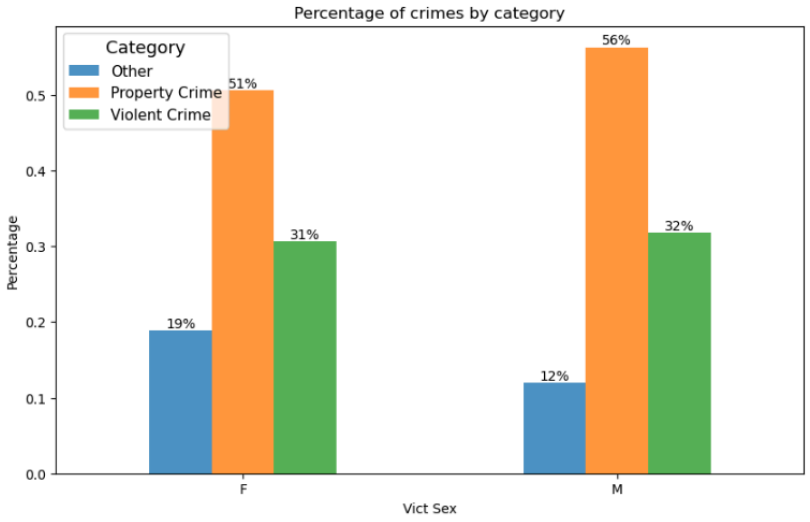


Figure 3 - Percentage of crimes by category and victim's age

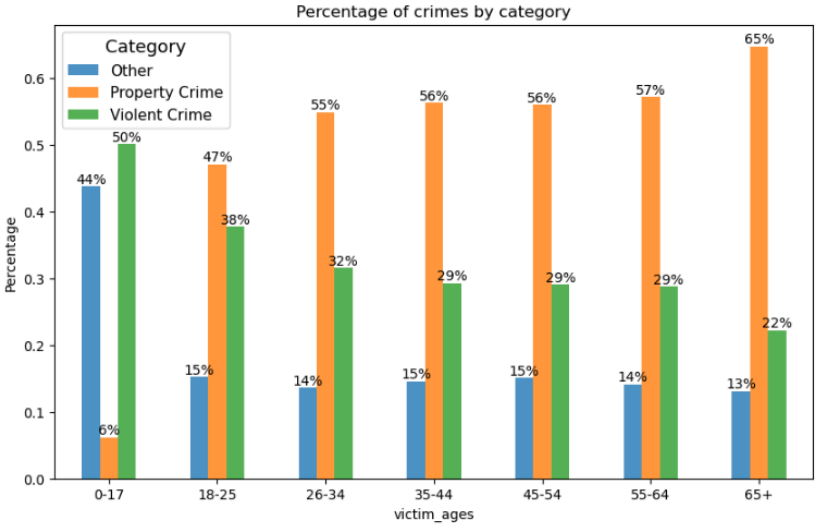


Figure 4 - Percentage of crimes by category and period

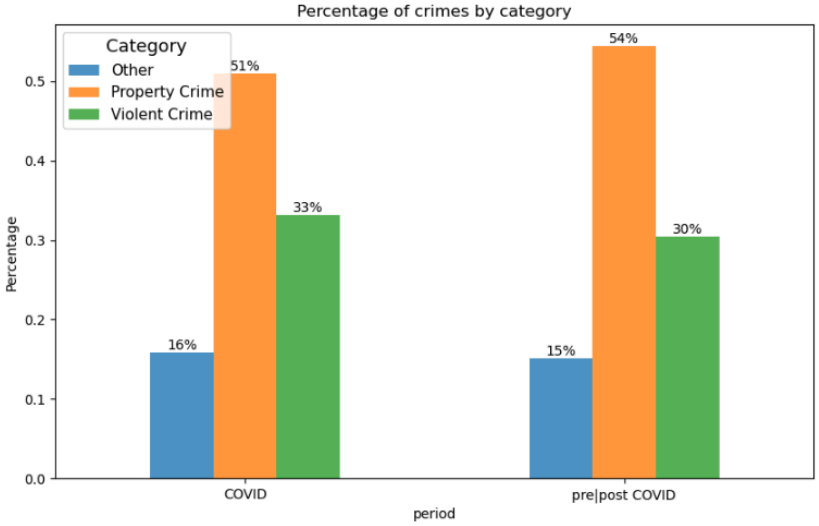


Figure 5 - Percentage of homicides by victim's sex

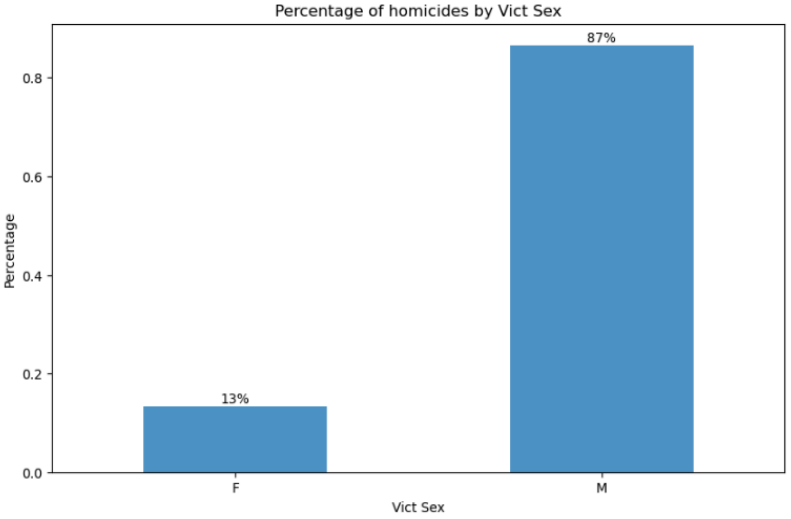


Figure 6 - Percentage of homicides by victim's age

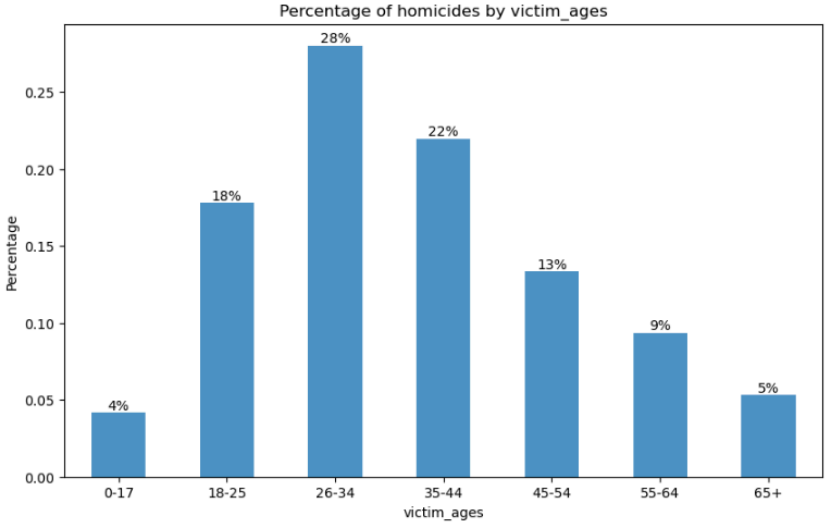


Figure 7 - Homicides by period over the dates

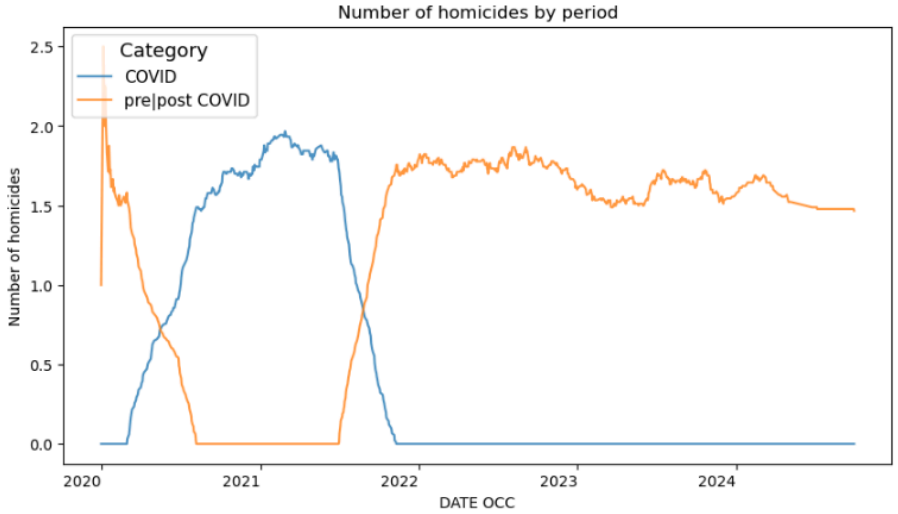


Figure 8 - Homicides by victim's sex over the dates

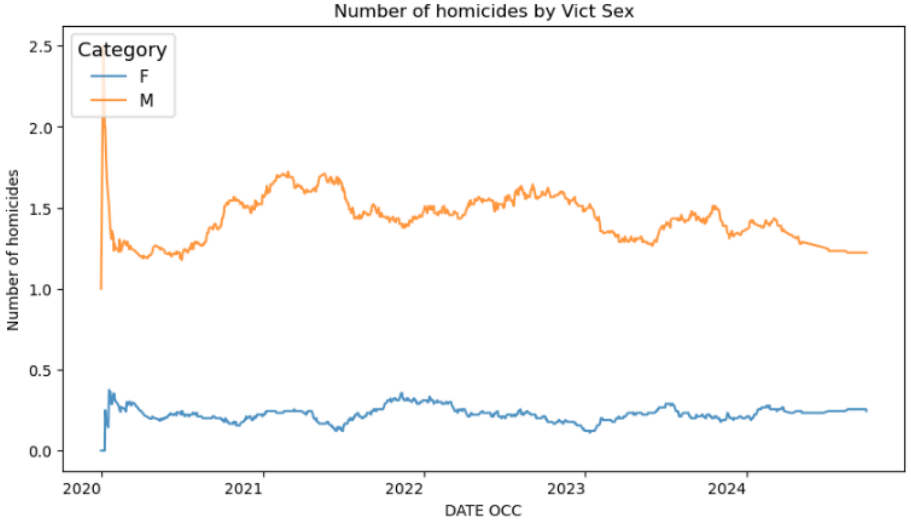
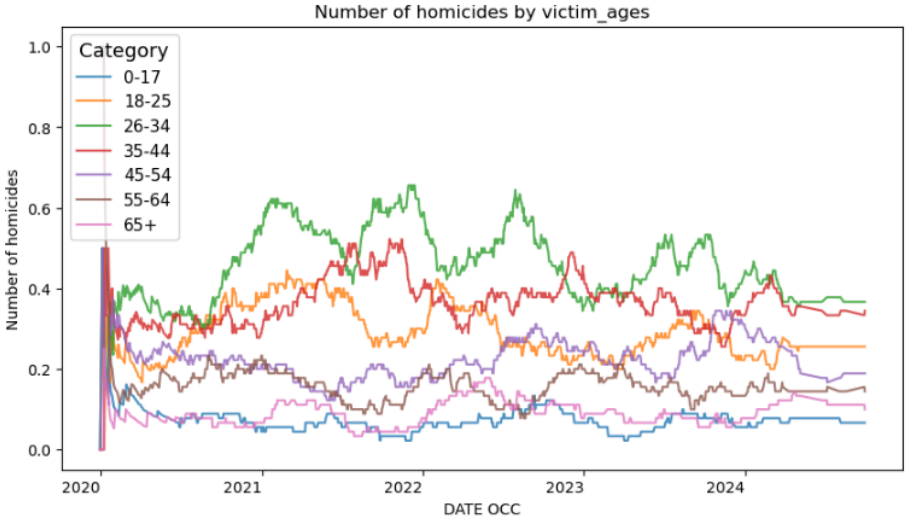


Figure 9 - Homicides by victim's age over the dates



We can conclude that the visualizations successfully achieved our expectations of them being:

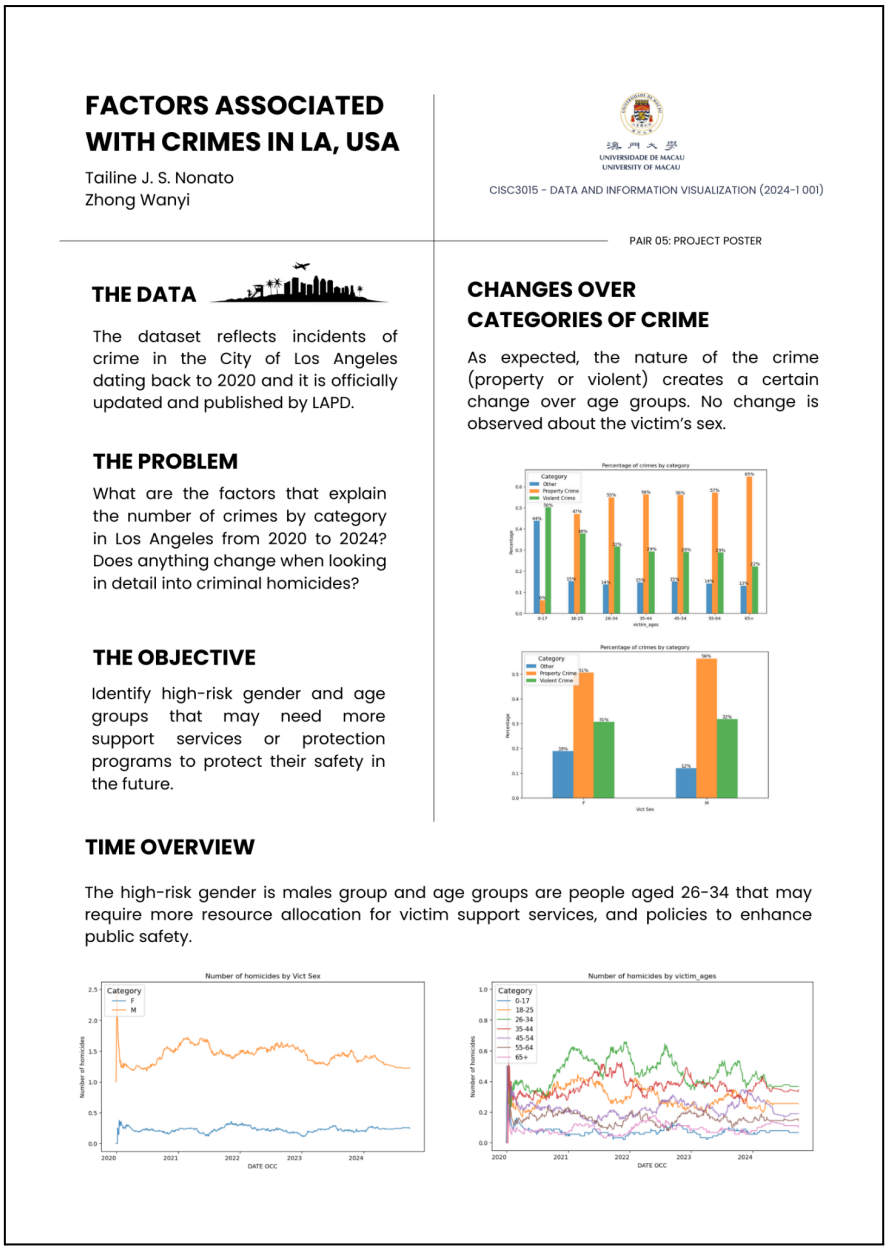
- truthful to the data, as none of them as deceiving from the original data in any way;
- functional, as all of them comply to their analytical purposes;
- beautiful, as all of them meet with standardized colors and shapes;
- insightful, as all of them offered valuable information to answer our research question.

V. Digital story

The digital story was produced with the help of Zoom, by doing a rerun of the Project Presentation, but updated with the final results and conclusions.

VI. Project poster

Figure 10 - Project poster



Project Limitations

(1) Time Limitations. We spend only one month to finish such a project, if we have more time and search for a more useful dataset, we can get a better result.
(2) The Influence of Covid-19 from 2020 to 2024.
(3) The Data of Crime Rate We Have not. Based on the dataset, we only can extract the number of incidents instead of the crime rate per 1,000 people, but if the dependent variable was crime rate, our conclusion would be more accurate and obvious.
(4) The Feature of Variables Is Discrete. Hence, we can not imply the method of correlation coefficient directly to find out their relationship.
(5) The Number of Unknown Gender Represented as 'X' Are Cleaned. But it is significant to affect the proportion of males and females in real situations.
(6) The Difficulty of Identifying and Affirming the relationships.

Project Learning

Firstly, we have become more adept at data visualization techniques and try our best to finish our pair project successfully. During the process of doing our project, we have learnt more about exploratory data analysis. Finally, we make an obvious conclusion to tackle our issue. According to the result of our project, as much as we did find visual indication that there is some sort of relation between crime category and victim's age, and between homicide and victim's sex, it would be hard to affirm these relationships based only on that. But we can know that the age group of 26-34 has the highest number of victims, with a significantly higher number of male victims compared to female victims. Hence, the high-risk gender is males group and age groups are people aged 26-34 that may require more resource allocation for victim support services, and policies to enhance public safety.

Project Division of Work

Table 1 - Division of Work

Task	**Responsibility**
Project presentation	Evy, Tailine
Coding	Evy, Tailine
Digital story	Evy
Poster	Tailine
Project report	Evy, Tailine
Forum Showcase	Evy
Organizing and zipping files	Tailine