Homicide Trends in Los Angeles Police Department Precincts

PSY6422 Module Project

2024-11-22

Introduction

Background

The COVID-19 pandemic significantly impacted society in a number of ways. These range from direct effects of contracting the illness to the mental health and economic impacts of related lockdown and social distancing measures implemented.

Despite initial assumptions that the disease did not discriminate, it soon became evident that individuals of certain demographic groups (ethnicity, age, income) were harder hit. In the US state of Michigan, despite only representing 14% of the population, Black residents accounted for 40% of COVID related deaths in early stages of the pandemic Vox, 2020.

Later, once a vaccine had been developed, serious inequalities in vaccine rollout between high income and low income individuals were uncovered. In the state of California, this was observed at a 60% difference in vaccination rates between wealthier vs. poorer areas Boston University, 2022.

In the UK, a report by the Equality and Human Rights Commission highlights a number of ways in which, without ever having contracted the illness, disadvantages in Britain have widened for individuals of protected characteristics such as gender and race.

With this in mind, I wanted to explore a topic relating to the pandemic. I chose to explore homicide rates as this seemed like an area that may see some change under the circumstances we saw in 2020. It seemed equally plausible to me for the change to be in the form of an overall increase or decrease, meaning it felt like a good question to explore with data.

Research Questions

- **How has the COVID-19 pandemic impacted on homicide rates in Los Angeles?**
- **If there was a change, was everyone equally likely to experience it?**

Data Origins

The project focuses on LAPD data due to the richness of location data associated with crime reports and the openness of available data. The LAPD is also one of the largest police departments in the USA allowing for a large amount of data to be explored in the project. Over 3 million rows of crime data were obtained for analysis.

Raw data for this project were obtained from the website sources listed below:

- 2010 - 2019 LAPD crime data
- 2020 - present LAPD crime data
- LAPD precincts geojson file

NB: The above datasets were extracted for analysis and visualisation in November 2024. The raw data in these datasets will continue to change in future. Any visualisations and insights drawn are exclusively related to the data which were available on the date of extraction

Setting Up

Loading Packages

The renv package was used in this project. Find the renv.lock file in my repository for more detailed information about the libraries used and their versions.

A brief list of the libraries installed and loaded for this project is also included below:

Package	Description
here	Easy management of file paths and working directory
tidyverse	For related packages to manage data
sf	For working with spatial data
gganimate	For animated plots
gifski	For animation rendering
renv	For managing packages

Loading Data

```
# csv crime data
rawdata2010s <- read_csv(here("data", "raw", "crime_rawdata_2010s.csv"))
rawdata2020s <- read_csv(here("data", "raw", "crime_rawdata_2020s.csv"))

# geojson spatial data
geo_map1 <- st_read(here("data", "map_data", "geo_data.geojson"))
```

```
## Reading layer ‘geo_data’ from data source
##   ‘/Users/paulo/Documents/MSc Psychological Research Methods with Data Science/Modules/PSY6422 Data A
##   using driver ‘GeoJSON’
## Simple feature collection with 21 features and 5 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: -118.6682 ymin: 33.70366 xmax: -118.1554 ymax: 34.33731
## Geodetic CRS:  WGS 84
```

Data Wrangling

Initial sanity checks

I started by visually checking over the data. A few things I checked for and noticed are noted below:

- I started by understanding what the data was describing and identifying columns of interest. This informed the first steps of my data wrangling below.
- I noticed the “DATE.OCC” column in the crime data had the same time (midnight) for every row which indicated to me that it is false information.

Processing crime data: Part 1

Informed by the initial checks, I began processing the crime data. I chose to merge the two crime datasets early on as they include the exact same variables, just covering different time spans. Following the merge I had over 3 million observations.

Due to the size of the dataset I prioritised trimming down to the relevant data for my project. This involved trimming unnecessary columns and crimes which were not of interest. I also changed the formatting of columns for ease of interpretability and consistency in any later code.

```
# changing DR_NO data class to character to match 2010s dataset and allow bind_rows to run
rawdata2020s_classchange <- rawdata2020s %>% mutate(DR_NO = as.character(DR_NO))

# merge the two separate datasets into one
crimedata1_merged <- bind_rows(rawdata2010s, rawdata2020s_classchange)

# trim unnecessary columns
crimedata2_coltrim <- crimedata1_merged %>%
  select(
    `DATE OCC`,
    `AREA`,
    `AREA NAME`,
    `Crm Cd Desc`
  )

# renaming remaining columns
crimedata3_colrename <- crimedata2_coltrim %>%
  rename(
    "date_occ" = `DATE OCC`,
    "precinct_num" = `AREA`,
    "precinct_name" = `AREA NAME`,
    "crime_type" = `Crm Cd Desc`
  )

# checking for crime types of interest for the project
unique(crimedata3_colrename$crime_type)

# creating homicide variable to filter for homicide only (lynching category was included
# in this as I aim to focus on intentional and unlawful killings).
homicide <- c(
    "CRIMINAL HOMICIDE",
    "LYNCHING"
    )

# filtering for homicide crimes only
homicidedata1 <- crimedata3_colrename %>%
  filter(
    crime_type %in% homicide
  )
```

Second round of sanity checks

After some initial filtering and reformatting to make the data more manageable, I went through another round of checks. Each line of code below has a preceding comment explaining what I was checking.

In addition to the below checks, I also completed a visual sanity check to ensure the precinct numbers on both spatial and crime datasets referred to the same area and found no issues.

```
# the data should cover a period from January 2010 - October 2024. Output does not show
# correct range. Error is due to R treating this column as a character variable. Date
# information will be extracted.
range(homicidedata1$date_occ) # "01/01/2010 12:00:00 AM" "12/31/2022 12:00:00 AM"


# following above check, now checking all data classes. Noted all will need to be changed.
# date_occ will be changed from "character" to "date", the remainder will be changed to
# "factor" as they are categorical variables.
columns <- c("date_occ", "precinct_num", "precinct_name", "crime_type")
sapply(homicidedata1[columns], class)

# there should only be 21 LAPD precincts. This is correct, but noticed geo_map will need
# padding to allow join with homicide data.
range(homicidedata1$precinct_num) # 01 - 21
range(geo_map1$PREC) # 1 - 21
```

Processing crime data: Part 2

I identified a few other things to address within the checks above. I went through a second round of processing the data based on these checks.

```
# changing categorical from character to factor
homicidedata2_factors <- homicidedata1 %>%
  mutate(
    `precinct_name` = as.factor(`precinct_name`),
    `crime_type` = as.factor(`crime_type`),
    `precinct_num` = as.factor(`precinct_num`)
  )

# changing date from character to date class following sanity checks above
homicidedata3_date <- homicidedata2_factors %>%
  mutate(
    date_occ = as.Date(sub(" .*", "", date_occ), format = "%m/%d/%Y")
  )

# checking range of dates - range is now as expected.
range(homicidedata3_date$date_occ) # "2010-01-01" "2024-10-07"

# creating summary data - not including 2024 data as we don't yet have the full year so
# won't be included in visualisations
homicide_summary_data <- homicidedata3_date %>%
  mutate(year = year(date_occ)) %>%
  filter(year < 2024) %>%
  group_by(precinct_num, precinct_name, year) %>%
  summarise(
    homicide_count = n(),
    .groups = "drop"
  )
```

Processing spatial data

Next, I focused on applying some changes to the spatial data. This was mostly just formatting and preparation for a join, specifically:

- I trimmed away all columns, keeping only the precinct_num column (for joining) and the geometry columns (containing polygons).
- I renamed the precinct_num column to set up for the join.
- I padded the precinct_num column as I identified this inconsistency between the two datasets in earlier checks.

```
# renaming and trimming columns
geo_map2_renamed <- geo_map1 %>%
  rename(
    "precinct_num" = `PREC`,
  ) %>%
  select(
    precinct_num,
    geometry
  )

# padding the precinct_num column
geo_map3_pad <- geo_map2_renamed %>%
  mutate(
    # changing precinct_num to character to allow padding
    precinct_num = as.character(precinct_num),
    # adding 0 to the left of any characters which aren't already a length of 2
    precinct_num = str_pad(
      geo_map2_renamed$precinct_num,
      width = 2,
      side = "left",
      pad = "0"
      ),
    # finally changing precinct_num to factor as it's a categorical variable
    precinct_num = as.factor(precinct_num)
  )
```

Joining the data

In the line of code below I am executing the join of spatial and crime data ready for visualisations later in the process. I also complete one final missing values check to identify any potential issues with the join but found no concerns. Note, I also visually reviewed the data for any inconsistencies following the join but found no issues.

```
# performing left join on data
joined_spatial_data <- left_join(geo_map3_pad,
                                 homicide_summary_data,
                                 by = "precinct_num") %>%
  arrange(year, precinct_num)

# checking for missing values
colSums(is.na(joined_spatial_data))
```

Saving the processed data

I chose to save the processed data (both joined spatial data and summary data) used in the subsequent visualisations. This will make it easier to pick up processed data and skip the data preparation code for those who prefer it.

```
# saving summary data
write_csv(homicide_summary_data, here("data", "processed", "homicide_summary_data.csv"))

# saving joined data
st_write(joined_spatial_data, here("data", "processed", "joined_spatial_data.geojson"))
```

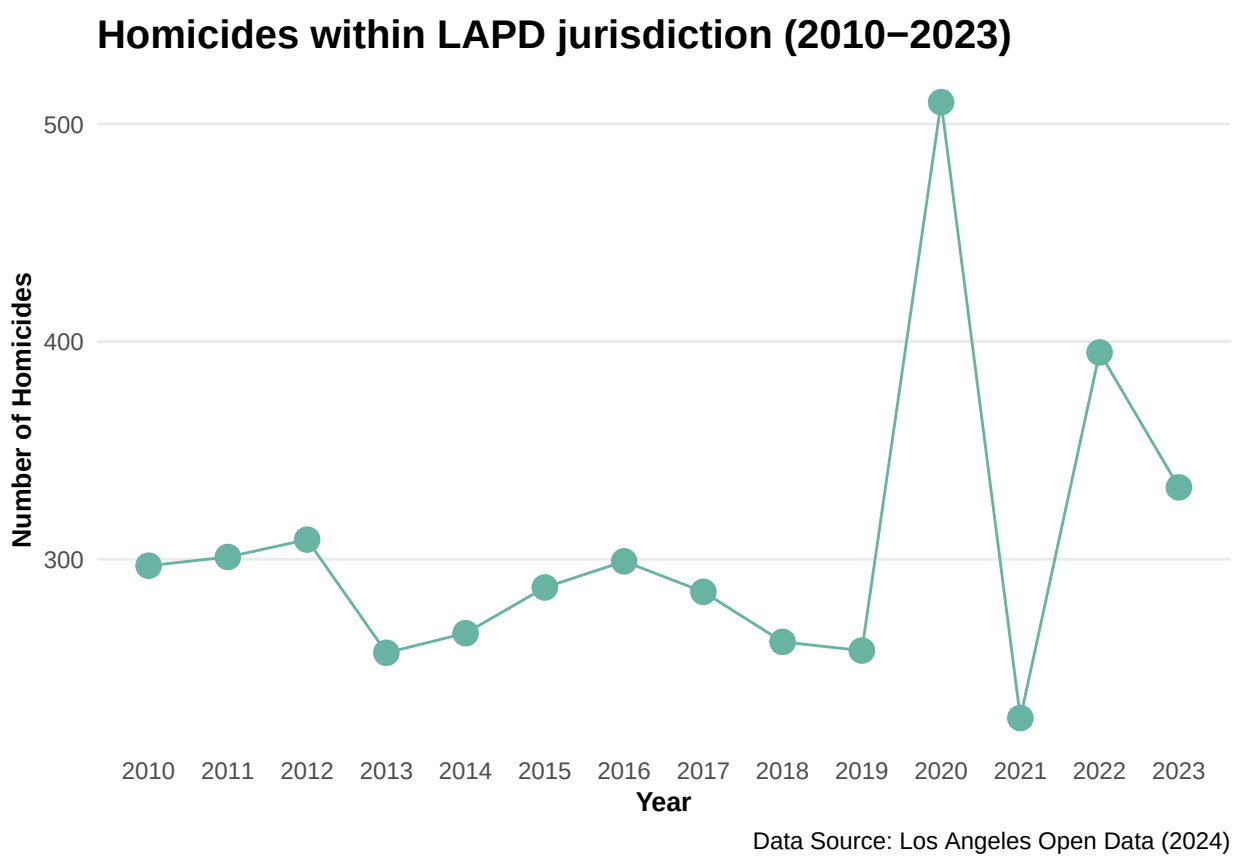
Data Visualisation

Visualisation 1: Connected scatterpot

I started the visualisation process by exploring the patterns in the data before determining my next steps. I started with a simple line graph and quickly noticed a considerable spike in homicides in 2020. I noticed a general decline in the years following 2020, but a very steep drop in 2021 seemed to stick out from the general trend downwards post-pandemic.

```
# creating a connected scatterplot
homicides_scatter <-
  # using summary data
  homicide_summary_data %>%
  # grouping by year to show overall summary across LAPD
  group_by(year) %>%
  # total count of homicides
  summarise(total_homicide_count = sum(homicide_count)) %>%
  # mapping aesthetics
  ggplot(aes(x = year, y = total_homicide_count)) +
  # adding simple line graph layer
  geom_line(colour = "#69b3a2") +
  # layering with scatterplot for connected scatter graph
  geom_point(color = "#69b3a2", size = 4) +
  labs(  # editing labels
    title = "Homicides within LAPD jurisdiction (2010-2023)",
    # citing the data source
    caption = "Data Source: Los Angeles Open Data (2024)",
    x = "Year",
    y = "Number of Homicides"
  ) +
   # opting for a minimal theme
  theme_minimal() +
  theme(
    # title font
    plot.title = element_text(size = 15, family = "Helvetica", face = "bold"),
     # axis title font
    axis.title = element_text(size = 10, family = "Helvetica", face = "bold"),
    # removing all minor gridlines
    panel.grid.minor = element_blank(),
    # removing major x gridlines
```

```
    panel.grid.major.x = element_blank()
  ) +
  # including all years on x axis
  scale_x_continuous(breaks = seq(2010, 2023, by = 1))

# viewing the plot
homicides_scatter
```



Saving the first visualisation

```
# saving the plot
ggsave(
  filename = here("plots", "homicides_scatter.png"),
  plot = homicides_scatter
  )
```

Visualisation 2: Animated choropleth map

Following my initial graph, I felt it would be helpful to visualise the trends geographically for a different perspective on the data.

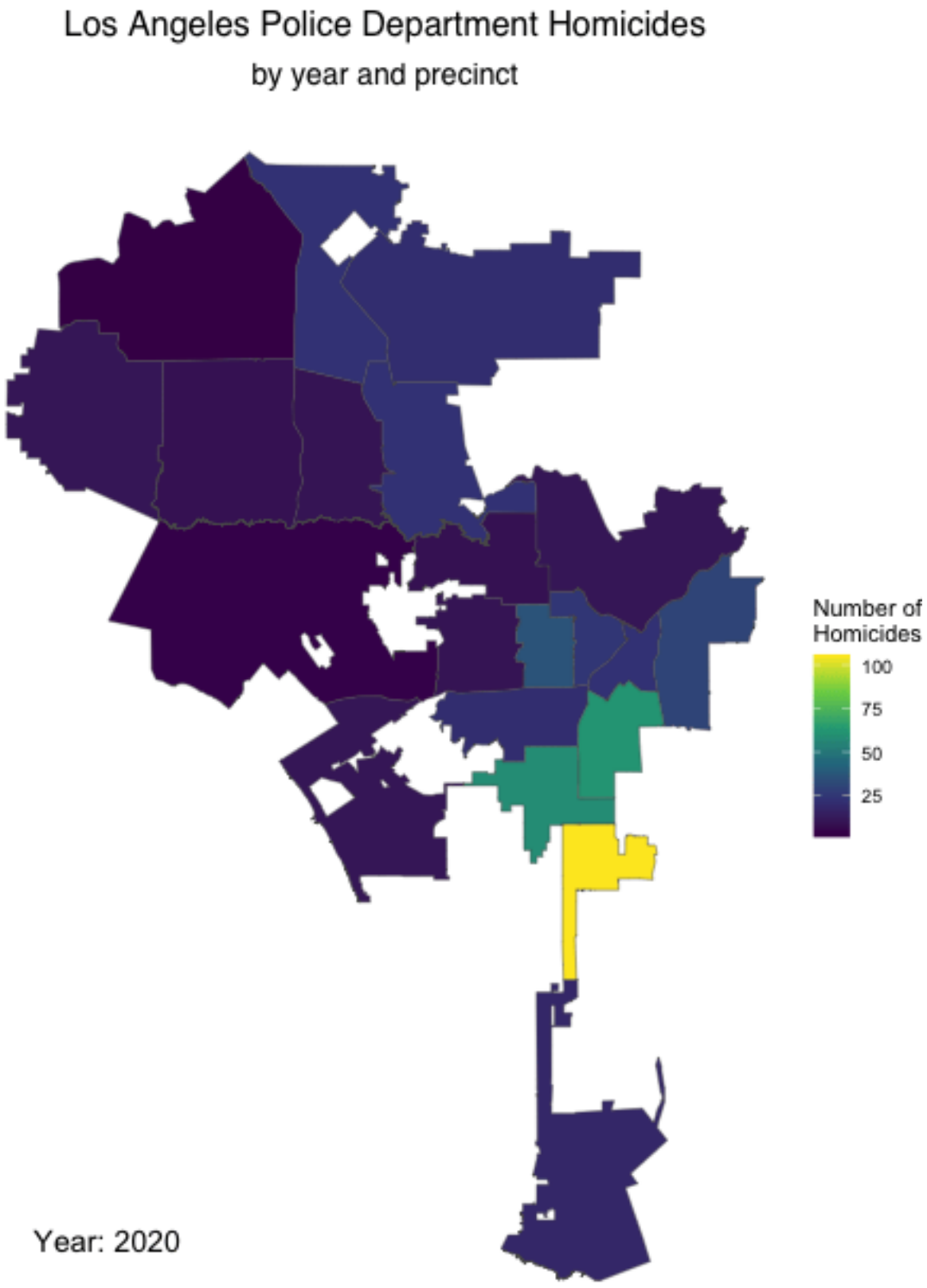
I found this visualisation particularly helpful in highlighting two main issues:

- First, I noticed a considerable lack of data in 2021 compared to all other years, with entire precincts lacking data for the whole year. This explained the steep decline in my first visualisation which seemed out of trend. It also raised a new issue with how best to visualise the data.
- Second, I noticed that, despite a considerable impact of the pandemic on homicides as a whole, the effect seems concentrated in certain parts of the map. This was my initial confirmation of the research question.

Following these insights I gained from my second visualisation, I began to consider how best to craft my final visualisation.

```
# creating the animation
homicides_choropleth <-
  animate(
    # using joined spatial dataset
    ggplot(joined_spatial_data) +
      # fill by number of homicides in a polygon each year
      geom_sf(aes(fill = homicide_count)) +
      geom_text(
        # dynamic year label
        aes(label = paste("Year:", year)),
        x = -118.65,
         # manual coordinates for label - chose bottom left due to layout of map
        y = 33.72,
        # adjusting format of label
        size = 5, color = "black", hjust = 0, vjust = 0, check_overlap = TRUE
      ) +
      # animating by year
      transition_time(year) +
      # minimal theme
      theme_void() +
      # colour-blind friendly scale applied
      scale_fill_viridis_c() +
      labs(# adding labels
        title = "Los Angeles Police Department Homicides",
        subtitle = "by year and precinct",
        fill = "Number of\nHomicides",
        caption = "Data Source: Los Angeles Geohub (2024); Los Angeles Open Data (2024)"
      ) +
      theme(
        # adjusting formatting of text in visualisation
        plot.title = element_text(size = 16, family = "Helvetica", hjust = 0.5),
        plot.subtitle = element_text(size = 14, family = "Helvetica", hjust = 0.5),
        plot.caption = element_text(face = "italic", size = 10)
      ),
    # setting number of years as number of frames
    nframes = length(unique(joined_spatial_data$year)),
    # setting width
    width = 450,
    # setting height
    height = 650,
    # low fps to allow time to look at each choropleth map shown
    fps = 1
  )
```

```r
# view the animation
homicides_choropleth

# saving the animation
anim_save(
  filename = here("plots", "homicides_choropleth.gif"),
  animation = homicides_choropleth
  )
```



Data Source: Los Angeles Geohub (2024); Los Angeles Open Data (2024)

Visualisation 3: Dumbbell Plot

Choosing the visualisation

For my third and final visualisation, I chose to focus on the steep jump in homicides between 2019 and 2020. I considered looking at "pre" and "post" pandemic averages but felt this would dilute the point of the visualisation which was to explore the spike in homicides and understand where this effect was most profound. Post-pandemic trends were also difficult to assess due to the considerable amount of missing data in 2021.

I wanted my visualisation to give the viewer a sense of "direction" of the effect. With 21 precincts, it was also important to not overload the viewer with too much visual clutter. I chose a dumbbell chart due to its ability to succinctly visualise two points for a number of categories and allow easy comparison between the two.

Formatting decisions

I tried to make purposeful decisions with my formatting to draw the viewer's attention to the story the visualisation is telling. One way I did this was by minimising grid lines to reduce background noise.

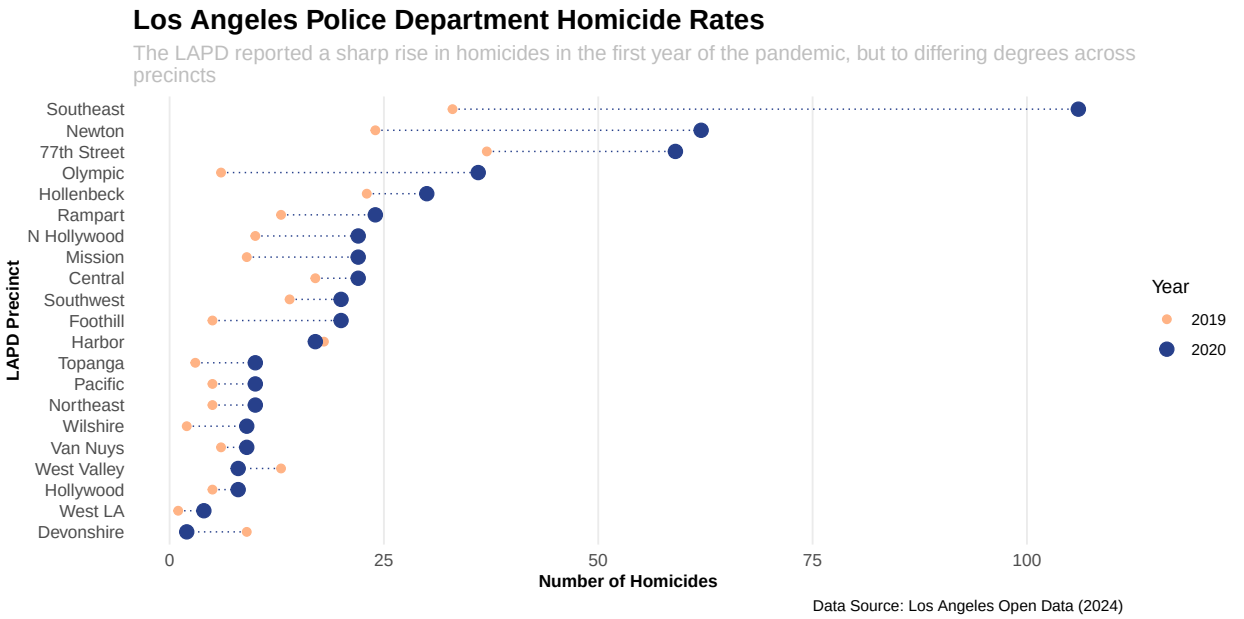
I also chose to colour and size the "2019" point more subtly, as this was to act as the "baseline" which gives scale to the change seen in 2020. I achieved this effect by giving the dotted line and the "2020" point the same colour. My aim in doing this was to give the viewer a sense of continuity between the two. My intention was to create an effect of distance travelled to "arrive" at the second data point.

Some other formatting decisions I made were in reducing text size of x/y labels and applying a grey colour to the subtitle. These choices were all made in an attempt to reduce the aspects of the graph pulling on the viewer's attention. These aspects are still visible for when the viewer wants that information, but they aren't the first thing the attention is pulled to.

This process did involve some re-wrangling of the data, so the code chunk starts with that.

```
# filtering and trimming away unnecessary data
dumbbell_data_filter <- homicide_summary_data %>%
  filter(year %in% c(2019, 2020))

# reformatting data
dumbbell_data_reformat <- dumbbell_data_filter %>%
  pivot_wider(
    names_from = year,
    values_from = homicide_count,
    names_prefix = "count_"
  )

# wrapping subtitle
wrapped_subtitle <- str_wrap(
  "The LAPD reported a sharp rise in homicides in the first year of the pandemic, but to
  differing degrees across precincts",
  width = 110)

# creating the plot
homicides_dumbbell <- ggplot(dumbbell_data_reformat) +
  # adding segment layer for connecting line
  geom_segment(aes(
    # setting start and end points of segments
    x = count_2019, xend = count_2020,
    y = reorder(precinct_name, count_2020)
  ),
```

```
  # creating dotted line - colour matches point of interest (2020 count)
  linetype = "dotted", color = "#27408B", linewidth = 0.4) +
  # layering a point illustrate 2019 data
  geom_point(
    # small orange dots for 2019
    aes(x = count_2019,
        y = reorder(precinct_name, count_2020),
        color = "2019"),
        size = 2) +
  # layering a point to illustrate 2020 data
  geom_point(
    # larger blue dots for 2020
    aes(x = count_2020,
        y = reorder(precinct_name, count_2020),
        color = "2020"),
        size = 3.5) +
  # setting point/legend colours
  scale_color_manual(
    name = "Year",
    values = c("2019" = "#FFB385", "2020" = "#27408B")
  ) +
  labs(
    # adding labels
    title = "Los Angeles Police Department Homicide Rates",
    subtitle = wrapped_subtitle, # inputting the wrapped subtitle from above
    caption = "Data Source: Los Angeles Open Data (2024)", # citing the data source
    x = "Number of Homicides",
    y = "LAPD Precinct"
  ) +
  # choosing a minimal theme
  theme_minimal() +
  theme(
    # y-axis label size and font
    axis.text.y = element_text(size = 10, family = "Helvetica"),
    # x-axis label size and font
    axis.text.x = element_text(size = 10, family = "Helvetica"),
    # title font
    plot.title = element_text(size = 16, family = "Helvetica", face = "bold"),
    # subtitle font
    plot.subtitle = element_text(size = 12, family = "Helvetica", color = "grey"),
    # axis title font
    axis.title = element_text(size = 10, family = "Helvetica", face = "bold"),
    # removing horizontal grid lines
    panel.grid.major.y = element_blank(),
    # removing minor vertical lines
    panel.grid.minor = element_blank()
  )

# viewing the final visualisation
homicides_dumbbell
```



Saving final visualisation and data

```
# saving final data
write_csv(dumbbell_data_reformat, here("data", "processed", "dumbbell_data_reformat.csv"))

# saving final visualisation
ggsave(
  filename = here("plots", "homicides_dumbbell.png"),
  plot = homicides_dumbbell
  )
```

Conclusion

Interpretation

My final visualisation clearly shows an increase in homicide rates from 2019 to 2020. The visualisation effectively breaks down the change by precinct, allowing the viewer to understand where the increases were most pronounced.

The four precincts with the highest rates of homicide in 2020 (Southeast, Newton, 77th Street, and Olympic), are all within the central and southeast areas covered by the LAPD. These areas are also the lowest income areas of Los Angeles Economic Roundtable.

It seems that the visualisation answers the initial questions. The COVID-19 pandemic resulted in a significant increase in homicide rates across Los Angeles. However, not for all residents. Residents of central and southeast Los Angeles were most likely to feel the impacts of this change. Incidentally, these are also some of the lower income areas of Los Angeles.

Reflections on the project

As this was my first time coding and working with data, there are a few reflections to discuss from my experience on this project.

Interpretability

First, an important improvement for my visualisation would be to make it more widely accessible and interpretable. I tried to make my visualisation with an "audience" in mind. I imagined communicating this information to local services, government, charities, etc. who may have stakes in understanding who the victims of homicide are and when/where people become more vulnerable. I considered the LAPD in this "audience" too, as the point of using their data was to understand trends at a local level where targeted changes and preparations could be made.

On reflection, in keeping such a specific audience in mind I also assumed a high level of pre-existing knowledge about Los Angeles' geography and neighbourhoods. The graph includes no more information about the areas aside from their names. That means that anyone trying to interpret the information who doesn't know these areas has to seek out detail elsewhere, rather than receiving the entire message from the visualisation. Including some information about the demographics of the precinct residents or homicide victims would add important context.

An additional point regarding interpretability should be made about the raw homicide count. Percentage change was initially considered but seemed to communicate the wrong message about the data, with small changes at times equating to inflated percentages if the homicide count in 2019 was already low. A relative population density figure (e.g. number of homicides per 1000 residents) would have been the ideal fix for this. It would allow for more direct comparison between precincts and even to other cities.

Availability of data

A limitation of this project was the availability of data. With 2021 data completely missing in almost half of the precincts and very poor recording in the remainder of precincts, including any data from this calendar year in a visualisation would have been misleading. That also meant exploring overall trends following the pandemic became considerably more difficult.

Breaking the data down further

My project looked at homicide rates as a whole but insights from the visualisations could be made more actionable by grouping in a few different ways. For example, checking for gang-related homicides, drug-related homicides, domestic homicides. Any patterns here would be helpful for the "audience" I mentioned above to be able to make effective preparations. For example, if all other types of homicide remain the same but drug-related homicides increase with lockdown measures, the LAPD might focus their preparations for another lockdown (or similar measure) very differently than if data was just communicated as an overall increase in homicides. Similarly, if domestic homicides accounted for most of that change, outreach programmes and relevant charities may benefit from understanding where the effects are most concentrated.