Progression of Crime over Time

PSY6422 Final Coursework Submission

2023-05-11

PSY6422 Project Motivation

I believe the great advantage of being a PSY6422 student is that you have a high level of control when it comes to the topic of the final project.

I am truly interested in crime and anti-social behaviour. Choosing a topic of my personal interest made me persevere even when I thought this project could not be completed.

Crime can have can have a negative impact on personal level but it can also affect the entire society. The levels of crime in a country are are reflection of its levels poverty, inequality, and other social issues.

Crime trends over the years need to be monitored so that evidence-based decisions can be made regarding law enforcement and criminal justice. Moreover, recording an incidence is important in order to offer adequate support to the victims and possibly identify vulnerable groups.

Research Questions

1. How have crime levels changed over time? Are there any specific types of crime that have increased or decreased over the years?
2. Which crime type has the highest and lowest number of recorded cases?

Data Origins

This data was retrieved from the Office of National Statistics website, under the subcategory Crime in England and Wales Appendix tables. The information I present here is all available on their website.

Data Collection

- The Crime Survey for England and Wales (CSEW) is a face-to-face victimisation survey that collects data on people's experiences with crime.The survey is conducted by Kantar Public on behalf of the Office for National Statistics, since 1981.
- This is data collected by the police on crimes reported to them, including details such as the type of crime, location, and demographic information about the offender and victim.

For more information on data collection for this version of the data head, please visit https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingseptember2022

Data Preparation

This project assumes you have the following libraries installed and running:

```
# ------------------- LIBRARIES ----------


#the following packages need to be installed

# install.packages("here")
# install.packages("tidyverse")
# install.packages("readxl")
# install.packages("ggplot2")
# install.packages("RColorBrewer")
# install.packages("plotly")
# install.packages("htmlwidgets")

#run the following packages

library(here)
library(tibble)
library(dplyr)
library(stringr)
library(tidyr)
library(readxl)
library(ggplot2)
library(RColorBrewer)
library(plotly)
library(htmlwidgets)
```

Note: It is possible to replace "tibble", "dplyr", "stringr", and "tidyr" for "tidyverse". However, this was causing an unicode character code that I was unable to fix with several debugging steps.

The UK ONS is an excellent website that publishes quantitative data on different topics. Therefore, it seemed like the ideal website to collect data from on Progression of Crime over Time. However, their data requires a lot of cleaning before it can be translated into a graph.

Throughout the course, we learnt to load `.csv` files, considering that the raw data is stored in an Excel `.xlsx` file with many sheets, I was tempted to simply open the Excel document and make the necessary changes there.

Yet, I was adamant to keep the original source untouched and in order to do this, I looked for a method that would allow me to select the Excel sheet with data that was relevant to this project.

The data frame from TABLE A1 looks like this:

```
head(df)
```

```
## # A tibble: 6 x 76
##   A1a: Trends in CSEW in~1 ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10
##   <chr>                    <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Offence group [note 4]   Jan ~ Jan ~ Jan ~ Jan ~ Jan ~ Jan ~ Jan ~ Jan ~ Apr ~
## 2 VIOLENCE                 2063  2014  2265  2900  3768  4464  3746  3375  2551
## 3 Violence with injury     1226  1208  1592  1778  2407  2622  2365  1974  1507
## 4 Wounding                  544   454   613   791   874   1077  940   728   709
## 5 Assault with minor inju~ 682   806   978   987   1533  1545  1425  1245  798
## 6 Violence without injury  838   806   674   1122  1360  1843  1381  1401  1043
## # i abbreviated name:
## #   1: ‘A1a: Trends in CSEW incidents of crime, adults aged 16 and over/households (1,000s)‘
## # i 66 more variables: ...11 <chr>, ...12 <chr>, ...13 <chr>, ...14 <chr>,
## #   ...15 <chr>, ...16 <chr>, ...17 <chr>, ...18 <chr>, ...19 <chr>,
## #   ...20 <chr>, ...21 <chr>, ...22 <chr>, ...23 <chr>, ...24 <chr>,
## #   ...25 <chr>, ...26 <chr>, ...27 <chr>, ...28 <chr>, ...29 <chr>,
## #   ...30 <lgl>, ...31 <lgl>, ...
```

This was not tidy data, and there were a couple of problems I was happy to fix in the code above:

1. there are several ‘NA’ columns;
2. the initial rows in df do not contain any data, that needed to be eliminated;

Now, the following steps of data wrangling were essential to ensure the the data was arranged in the correct format, keeping only the main crime groups that were relevant to the graph. I wanted to make sure the column names reflected the data they contained to improve readability. Also, Table A1 contained two tables so we had to select the table that interested us “Trends in CSEW incidents of crime, adults aged 16 and over/households”.

```
# ------------------- DATA WRANGLE ----------


# select only the columns that include figures from April to March

df2 <- df %>%
  select ("A1a: Trends in CSEW incidents of crime, adults aged 16 and over/households (1,000s)",...10:.



# ensure the column names reflect the data in each column

colnames(df2) <- df2[1, ]


# keep only the main offence groups and omit the subgroups

df2 <- df2 %>%
  slice(2,7,8,38,47,52)


# transpose the data

df2 <- t(df2)


# again: ensure the column names reflect the data in each column

colnames(df2) <- df2[1, ]
```

```
# remove first row that is no longer needed

df2 <- df2[- 1, ]

# ensure column names are not written in all capitals

df2 <- as.data.frame(df2) %>%
  setNames(c("Violence", "Robbery", "Theft Offences",
             "Criminal Damage", "Fraud", "Computer Misuse"))



# create a new column called "year"
  #that is an extraction of the 5th to 9th characters in the rows


df2 <- df2 %>%
  mutate(year = str_sub(rownames(df2), start = 5, end =9)) %>%
  rownames_to_column(var = "row") %>% # name the row names column "row"
  select(-row) # delete this now unnecessary column
```

At this point, the data was looking tidier than when we first started. However, there were a few further steps that were specific so that the data could be translated into the ggplot.

```
# convert data from wide format to long format

df3 <- df2 %>%
  gather(key = colnames, value = "recorded_cases", -year) %>%
  rename(crime_group = colnames) # change "colnames" to "crime_group

# convert NA values into 0 so that NA value appear at the bottom of the graph

df3[is.na(df3)] <- 0


# check if y (recorded_cases) is a numeric value

if(is.numeric(df3$recorded_cases)) {
  print("TRUE")
} else {
  print("FALSE")
}
```

```
## [1] "FALSE"
```

```
# convert "recorded_cases" to numeric variable

df3$recorded_cases <- as.numeric(as.character(df3$recorded_cases))


# again: check if y (recorded_cases) is a numeric value

if(is.numeric(df3$recorded_cases)) {
  print("TRUE")
} else {
  print("FALSE")
}
```

```
## [1] "TRUE"
```

```
# show the the first few rows of processed data

head(df3)
```

```
##   year crime_group recorded_cases
## 1 2001    Violence           2551
## 2 2002    Violence           2579
## 3 2003    Violence           2355
## 4 2004    Violence           2191
## 5 2005    Violence           2221
## 6 2006    Violence           2287
```

Visualisations

Colour selection

You will see that I added a step to select a colour-blind friendly.

Since I was planning to build a graph that relied on colour to distinguish the different crime groups, making it accessible to the widest possible audience was important.

As someone whose colour blindness runs in the family, I understand that some people find colour communication difficult and struggle to understand the information presented. This is particularly problematic in scientific and medical research where accurate interpretation of data is critical.

Graph Selection

Different types of graphs could be used to represent the progress of crime throughout the years. Two of my main options were:

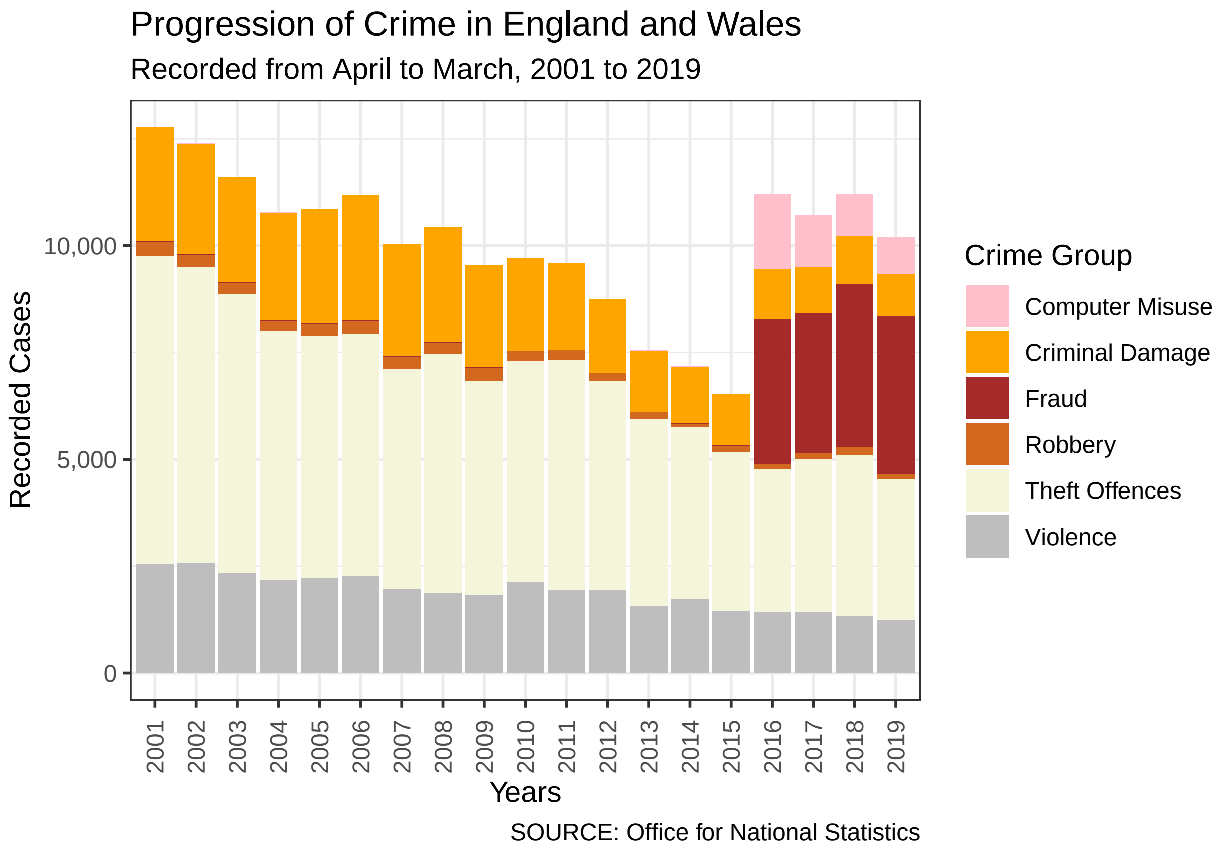
1. *a stacked bar chart*: that could be used to compare the relative proportion of different types of crime over time. Each year would be represented by a bar, and the year would be divided into segments representing the different types of crime.

2. *a line chart*: a simpler and more effective way to visualise trends in crime over time. Each year would be plotted on the x-axis, and the number of recorded cases for each crime would be plotted on the y-axis. Different colour lines would be used to represent different types of crime.

The stacked bar chart

```
b <- ggplot(df3, aes(x = year, y = recorded_cases, fill = crime_group)) +
  geom_bar(stat = "identity") +
  labs(title = "Progression of Crime in England and Wales", #plot a title
       subtitle = "Recorded from April to March, 2001 to 2019", #plot a subtitle
       x ="Years", y = "Recorded Cases", # plot the axes' titles
       fill = "Crime Group", # plot the colour legend title
       caption = "SOURCE: Office for National Statistics") +
  theme_bw() + #classic black and white look +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_fill_manual(values = c("pink", "orange", "brown", "chocolate", "beige","gray")) +
  scale_y_continuous(labels = scales::comma_format()) #format the y axis

# show the graph
b
```

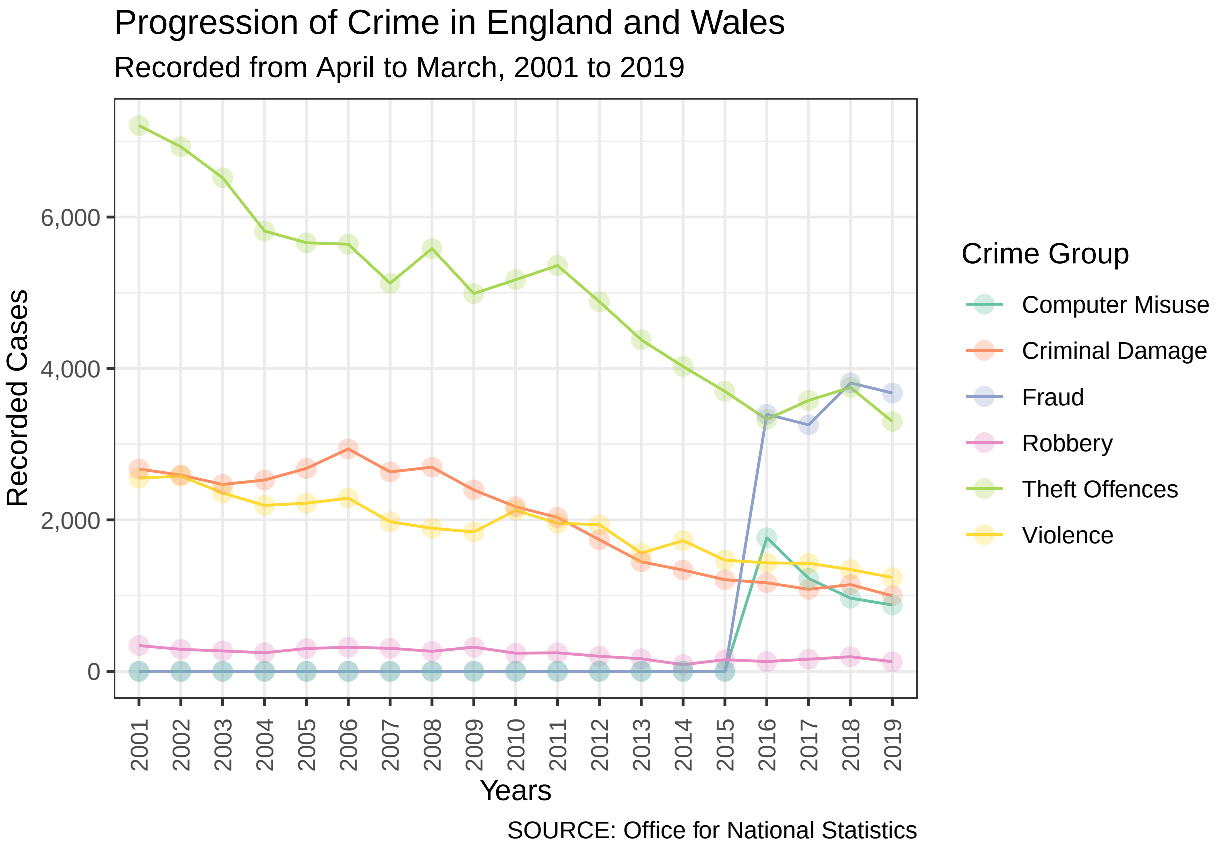


The line chart

```
#  to choose a pallet that is colour blind friendly: display.brewer.all(colorblindFriendly = TRUE)

p <- df3 %>%
  ggplot(mapping=aes(x=year, y=recorded_cases, group= crime_group,
                     colour = crime_group)) +
  geom_line() + #plot a line
  geom_point(size = 3, alpha = 0.3) + #plot the points
  labs(title = "Progression of Crime in England and Wales", #plot a title
       subtitle = "Recorded from April to March, 2001 to 2019", #plot a subtitle
       x ="Years", y = "Recorded Cases", # plot the axes' titles
       colour = "Crime Group", # plot the colour legend title
       caption = "SOURCE: Office for National Statistics") + #plot caption
  theme_bw() + #classic black and white look
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  # rotate the years 90 degrees
  scale_color_brewer(palette ="Set2") +
  #choose a pallet that is colour blind friendly
  scale_y_continuous(labels = scales::comma_format()) # format the y axis scale

# show the graph
p
```



As the main visualisation piece of this project, I chose the line chart. It was more aesthetically pleasing but more importantly it gives us immediate answers to the research questions.

We can immediately identify the overall trend and track the changes in data points from 2001 to 2019. Overall, it looks like crime has been decreasing over time or at least the number of recorded cases has decreased.

I argue that we want important information to reach and be accessible to a wide range of audiences. Therefore, creating a graph that is easy to read is an advantage. The lines make it easy to see the changes in data and points indicate the number of recorded cases.

Further, it is possible to compare how different crime groups have progressed compared to each other. We see that until 2015, 'theft offences' always had the highest number of recorded cases, but in 2016, 2018, and 2019 'fraud' was the crime group with the highest number of recorded cases. On the other hand, 'robbery' has consistently been the crime group with the lowest number of recorded cases.

One might think that 'fraud' and 'computer misuse' had zero occurrences. But it is important to remember that NA values were transformed into 0 so that they could be represented at the bottom of the graph. In reality, NA values mean that the data was not collected for that crime during that year. This is a limitation of the graph that is worth acknowledging as it could lead to misinterpretation.

Nevertheless, I still discarded the stacked bar chart. Since the stacks were of different sizes it created a number of problems that made the interpretation of the graph quite difficult and misleading. We could not easily compare different crime groups even with the addition of colour and it had limited ability to show the changes of each over time.

The final and interactive version

Now that I had selected my preferred graph, I wanted to make it interactive in order to allow the viewer to hover over each point and see more information about it. The y-axis is scaled, so once interaction was possible, the viewer could see the exact number of recorded cases for that specific type of crime in that year.

Also, it solved the problem that some points from different crime groups were close to each other and therefore, difficult to see which is highest.

I believe it was the right choice as it improved user engagement. Giving people the option to filter and focus on a specific subset of their interests. It is almost like having 6 graphs in 1.

```
# the chosen graph with hover over animation

p2 <- ggplotly(p)

p2 <- p2 %>%
  layout(title = list(text = "<b>Crime Progression Over Time</b> <br> <span style='font-size:14px'>
                      Recorded from April to March, 2001 to 2019 </span>",
                      x = 0.5), #add title and subtitle
             annotations = list(
               list(x = 1, y = 0, text = "SOURCE: Office for National Statistics",
                    showarrow = FALSE, font = list(size = 8),
                    xref = "paper", yref="paper", align="left", xanchor="left", yanchor="bottom", pad=li
             )) #add caption
```

Summary and Future Research

This visualisation project aimed to represent how crime progressed from 2001 to 2019. Moreover, it shows the different progression for the six main crime groups recorded by the ONS.

The graph leaves the overall impression that crime has been decreasing in England and Wales over the last few years. This is consistent with the statement published by the ONS in regard to this data.

While 'robbery' seems to have a lower number throughout the years, 'theft offences' that traditionally had the highest recorded cases have significantly decreased.

The ONS emphasises that the improvement in recorded processing and inclusion of new offences and other factors have made a contribution to the rise in recorded crime in some groups. For example, this can be seen in 'fraud' and 'computer misuse'.

Considering that it is known that the national lockdowns and social distance restrictions caused the number of recorded cases to fall significantly, for future research I would to understand the other factors driving this decrease in crime. It would interesting and useful to investigate a possible correlation between growth in income or the ageing population. This could then be taken further and compared with other geographical regions.

The study can also be a stepping stone into investigating the overall trend over the years and comparing it to the trends seen in possibly targeted groups based on their race, gender, religion, sexual orientation, and other factors for which they could be discriminated against. This would help create strategies to defend vulnerable groups in society.

Thank you for your time!