Lab 3: Effective Data Visualisations

The purpose of this lab is to practice identifying problems with data visualisations and to experiment with improvements based on established guidelines for more effective data visualisations.

NOTE that for some questions there may not be a unique correct answer; what is important is that you can provide a reason for what you are doing, based on the concepts and terminology used in the readings for this topic.

The Data Set

The data set is a CSV file, `nzpolice-proceedings.csv`, which was derived from "Dataset 5" of Proceedings (offender demographics) on the policedata.nz web site.

We can read the data into an R data frame with `read.csv()`.

```
crime <- read.csv("nzpolice-proceedings.csv")
head(crime)
```

```
  Age.Lower Police.District                                                ANZSOC.Division
1        15          Tasman                                  Acts Intended to Cause Injury
2        20   Auckland City Abduction, Harassment and Other Related Offences Against a Person
3        40   Auckland City Abduction, Harassment and Other Related Offences Against a Person
4        10   Auckland City                                  Acts Intended to Cause Injury
5        15   Auckland City                                  Acts Intended to Cause Injury
6        15   Auckland City                                  Acts Intended to Cause Injury
     SEX       Date
1 Female 2015-12-01
2 Female 2015-12-01
3 Female 2015-12-01
4 Female 2015-12-01
5 Female 2015-12-01
6 Female 2015-12-01
```

The following code reorders the levels of the `ANZSOC.Division` factor according to the highest age group count for each type of crime. It also generates `newlabels`, which are line-wrapped versions of the `ANZSOC.Division` levels.

```
types <- apply(table(crime$ANZSOC.Division, crime$Age.Lower), 1, max)
newlevels <- names(types)[order(types, decreasing=TRUE)]
newlabels <- unlist(lapply(strwrap(newlevels, width=30, simplify=FALSE),
```

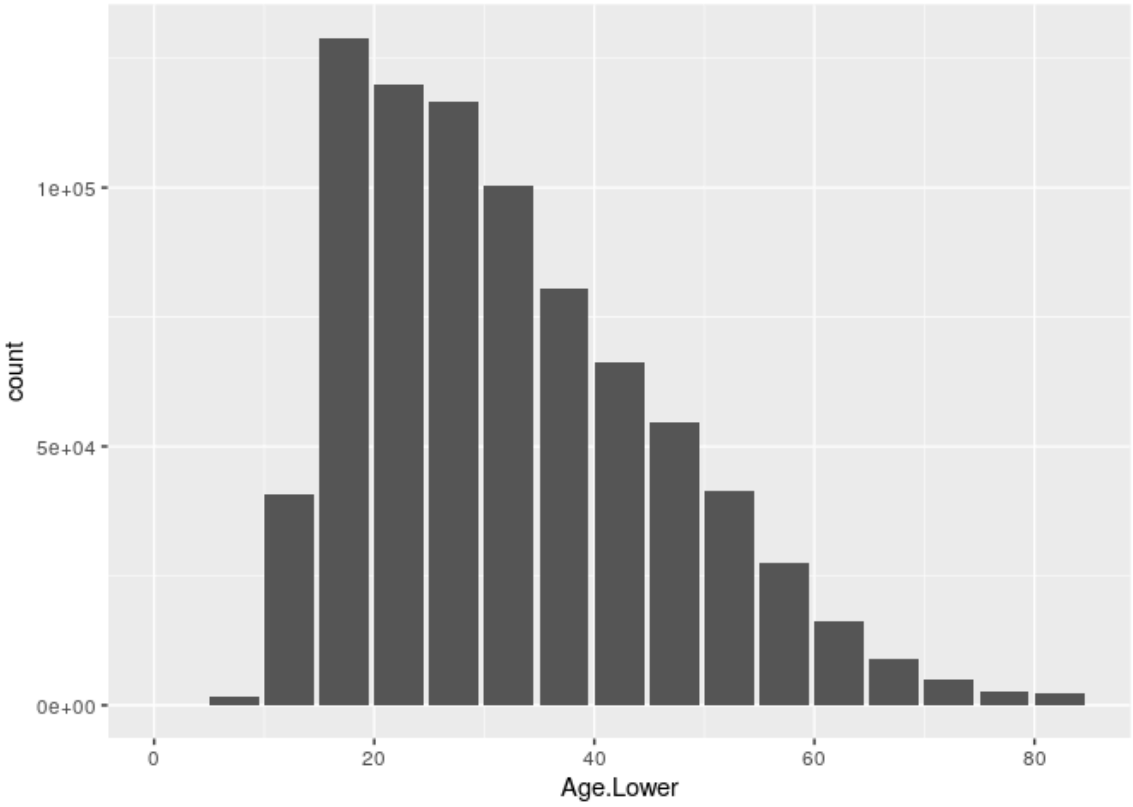
```
                            function(x) {
                                if (length(x) < 3)
                                    x <- c(x, rep(" ", 3 - length(x)))
                                paste(x, collapse="\n")
                            }))
crime$ANZSOC.Division <- factor(crime$ANZSOC.Division, levels=newlevels)
```

The following code generates a table of counts for the number of incidents per age group, broken down by type of crime.

```
crimeAgeType <- as.data.frame(table(crime$Age.Lower, crime$ANZSOC.Division))
crimeAgeType$Age <- as.numeric(as.character(crimeAgeType$Var1))
```

Questions of Interest

We have already seen the distribution of incidents across age groups (across all crimes); there is a highly skewed distribution with a peak of crimes in the 15-20 age group.



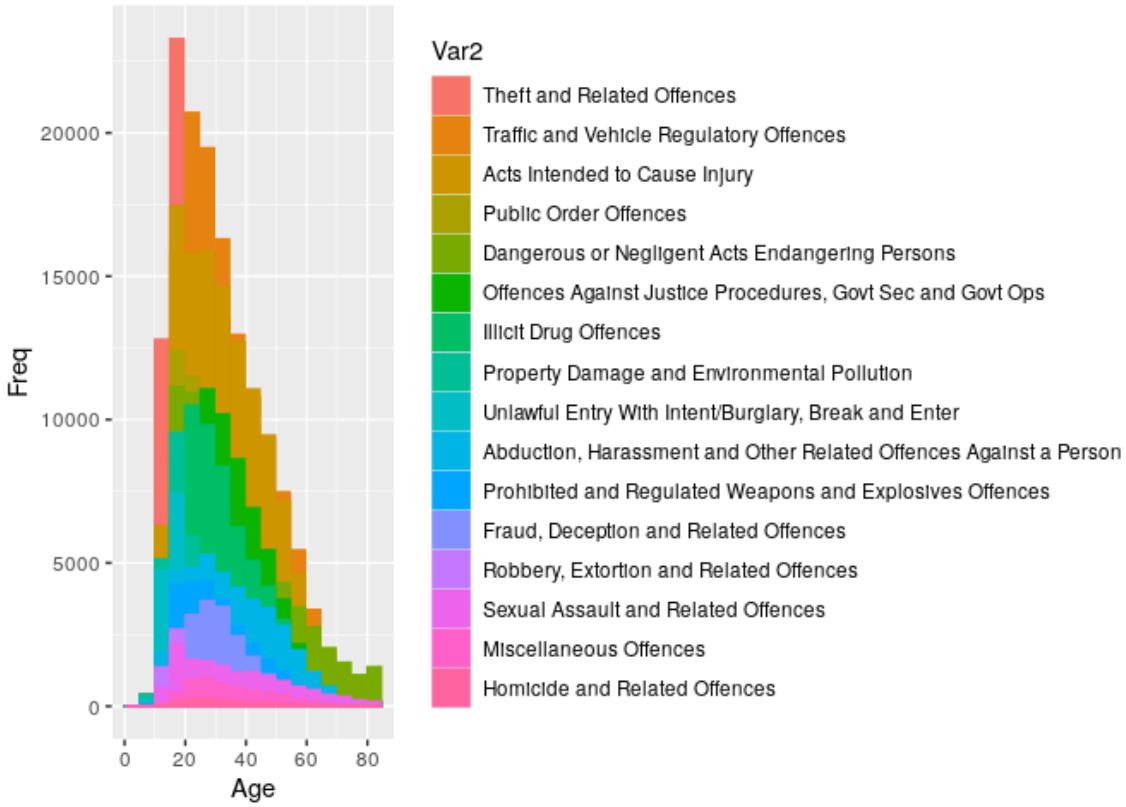
In this lab we will focus on the distribution of incidents across age groups, broken down by type of crime (`ANZSOC.Division`):

- What types of crime are the most common?
- Is the distribution of incidents across age groups the same for different types of crime?
  - Is there a single peak in all cases?
  - Is the peak in the same age group in all cases?

Data Visualisations

1. Run the following code to produce a bar plot of the number of incidents in each age group broken down by the type of crime.

```
ggplot(crimeAgeType) +
    geom_col(aes(Age, Freq, color=Var2, fill=Var2), position="identity",
             just=0)
```



Comment on what this data visualisation tells us about the questions of interest.

Comment on each of the following issues with this data visualisation:

- Overplotting.
- Colour use.
- Text labels.
- The principal of proportional ink.

You should write at least one sentence for each issue: is there a problem? how is the problem affecting our ability to answer the questions of interest?

2. **Write R code** to produce three modifications of the data visualisation from Question 1 that addresses overplotting by using each of the following techniques (one at a time):

- Semitransparency.
- Jittering (or, more generally, the `position` of the bars).
- Changing to a different geom (line or area).

Describe what changes you have made in each case and **comment** on whether you have improved the data visualisation in each case (is it easier or harder to answer the questions of interest?).

Do NOT use facetting in this question.

3. **Write R code** to produce a modification of the data visualisation from Question 1 that attempts to improve the labelling.

   Describe each change that you have made and **comment** on whether you have improved the data visualisation (is it easier or harder to answer the questions of interest?).

   Do NOT use facetting in this question.

4. **Write R code** to produce a modification of the data visualisation from the previous question that uses **small multiples**.

   Comment on whether this is an improvement on the previous data visualisation (is it easier or harder to answer the questions of interest?).

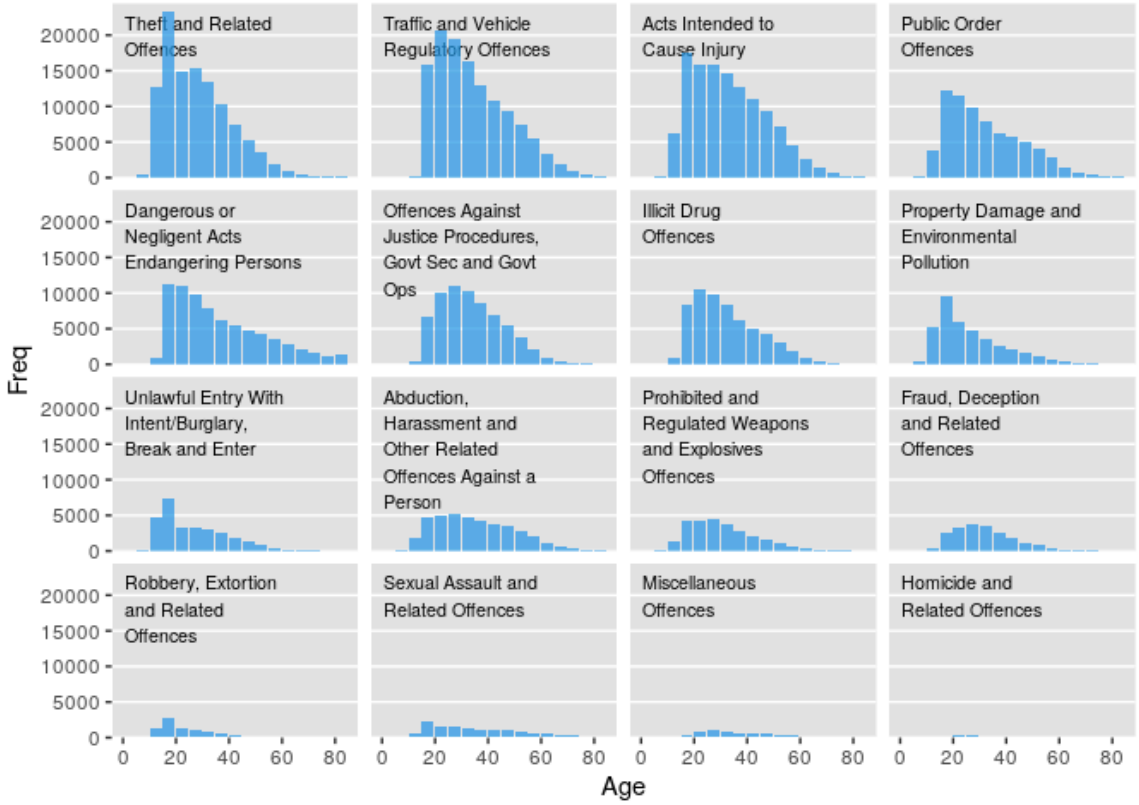
5. **Write R code** to produce a modification of the data visualisation from the previous question that attempts to increase the **data-ink ratio**.

   Describe each change that you have made and **comment** on whether you have improved the data visualisation (is it easier or harder to answer the questions of interest?).

6. **Write R code** that uses 'grid' (in combination with 'ggplot2') to produce the data visualisation below.

   NOTE that the `ANZSOC.Division` labels are top-left justified 2mm in from the top-left corner of each panel.

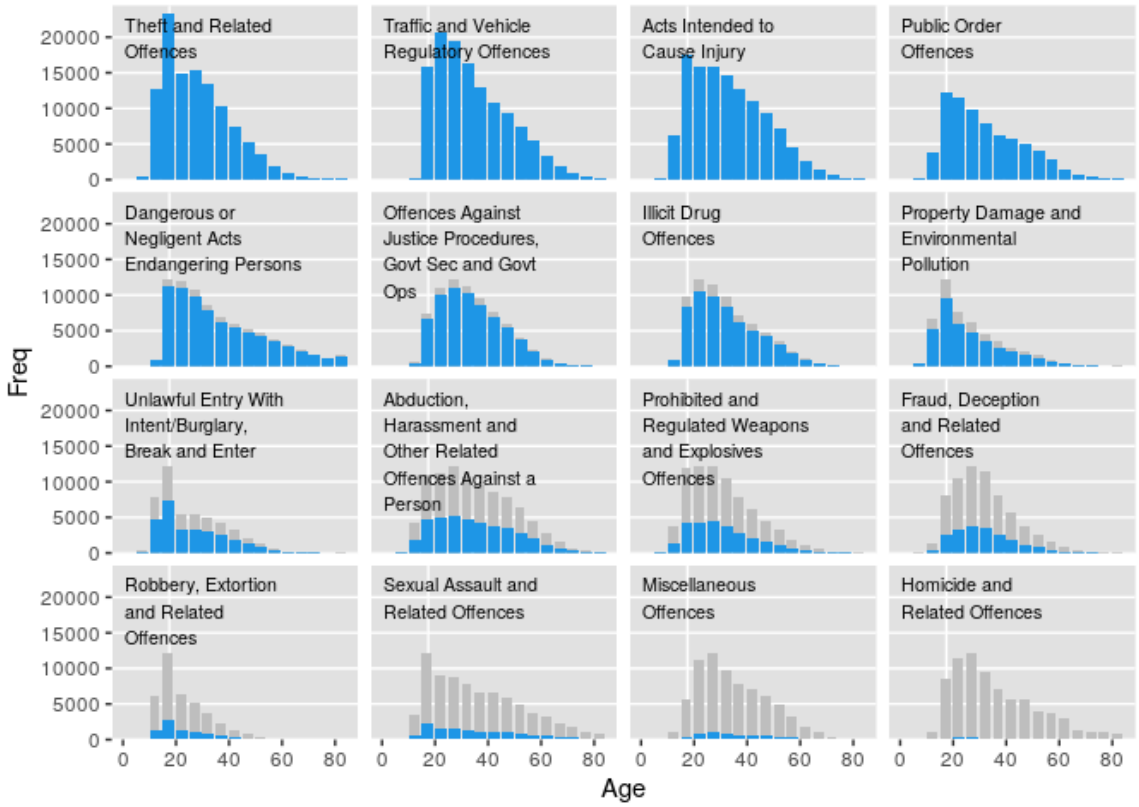
   Comment on whether this plot makes it easier or harder to answer the questions of interest.



Challenge

7. Can you produce the data visualisation shown below? Does this help with answering the questions of interest at all?

   NOTE that there is a vertical grid line in the middle of the 15-20 age group and the grey bars in each panel are the blue bars scaled up (or down) so that the highest bar is 0.5 the height of the panel.



The Report

Your submission should consist of a knitted R Markdown document, in HTML format, submitted via Canvas.

Your report should include:

- A brief description of the data and the question we are trying to answer.
- For each data visualisation, R code AND a brief text commentary.
- A brief overall summary.

Don't forget to also complete the Canvas Quiz!

Marking

Marks will be lost for:

- Plagiarism.
- Section of the report is missing.
- The summary is too short or does not make sense.
- Significantly poor R (or other) code.
- Overly verbose code, output, or commentary.

This work is licensed under a Creative Commons Attribution 4.0 International License.