```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

## Load the datasets

```
crime_raw <- read.csv("../../data/original/Crime_Data.csv")
force <- read.csv("../../data/original/Police_Use_of_Force.csv")
incidents_2021 <- read.csv("../../data/original/Police Incidents/Police_Incidents_2021.csv")
incidents_2022 <- read.csv("../../data/original/Police Incidents/Police_Incidents_2022.csv")
```

## Removing some obviously invalid parts of the data

```
crime <- crime_raw %>%
    filter(X != 0 & Y != 0)
```

## Understanding setdiff

setdiff returns the elements in the first set that does not appear in the second one

```
setdiff(c(1, 2, 3), c(2, 3, 4))
```

```
## [1] 1
```

## Why compare both use of force data and incident data to crime data

Crime data puts case numbers in two formats (two variables) with one format matching the use of force data and one format matching the crime data.

## Comparing crime and fuse of orce data

88.6% of case numbers in crime data can be found in use of force data.

```
length(setdiff(crime$Case_Number, force$CaseNumber)) / length(crime$Case_Number)
```

```
## [1] 0.886368
```

39.1% of case numbers in use of force data cna be found in crime data.

```
length(setdiff(force$CaseNumber, crime$Case_Number)) / length(force$CaseNumber)
```

```
## [1] 0.3910767
```

## Comparing crime data and incidents data

```
incidents_two_year_caseNumber <- c(incidents_2021$caseNumber, incidents_2022$caseNumber)
length(setdiff(incidents_two_year_caseNumber, crime$Case_NumberAlt)) / length(incidents_two_year_caseNu
```

```
## [1] 0.4616476
```

```
length(setdiff(crime$Case_NumberAlt, incidents_two_year_caseNumber)) / length(crime$Case_NumberAlt)
```

```
## [1] 0.6235916
```