```
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
```

```
library(lubridate)
```

```
##
## Attaching package: ’lubridate’

## The following objects are masked from ’package:base’:
##
##     date, intersect, setdiff, union
```

Load the datasets

```
crime_raw <- read.csv("../../data/original/Crime_Data.csv")
crime <- crime_raw %>%
    filter(X != 0 & Y != 0)
crime$Reported_Date <- ymd_hms(crime$Reported_Date)

force_raw <- read.csv("../../data/original/Police_Use_of_Force.csv")
force <- force_raw %>%
    filter(X != 0 & Y != 0)
force$ResponseDate <- ymd_hms(force$ResponseDate)

incidents_2019 <- read.csv("../../data/original/Police Incidents/Police_Incidents_2019.csv")
incidents_2020 <- read.csv("../../data/original/Police Incidents/Police_Incidents_2020.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")

dim(crime)
```

```
## [1] 208040     24
```

```
dim(na.omit(crime))
```

```
## [1] 207743     24
```

```
dim(force)
```

```
## [1] 36804    30
```

```
dim(na.omit(force))
```

```
## [1] 34749    30
```

```
dim(incidents_2019)
```

```
## [1] 23232    23
```

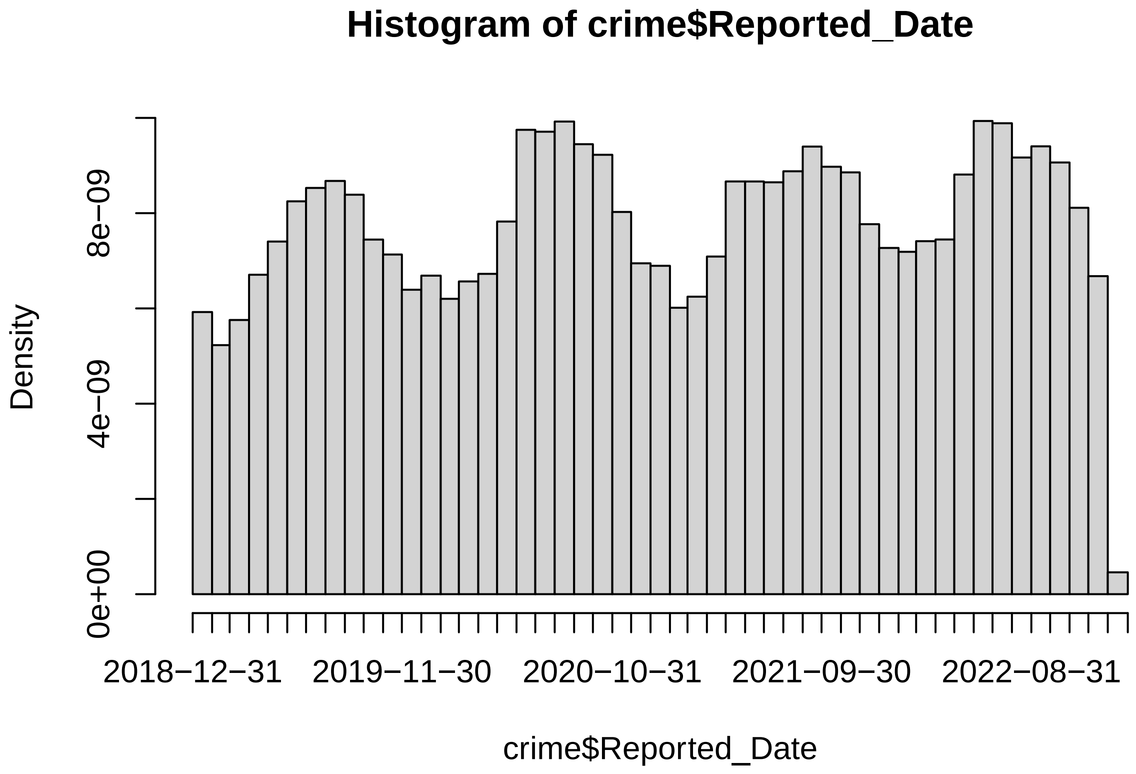
```
min(crime$Reported_Date)
```

```
## [1] "2019-01-01 00:01:58 UTC"
```

```
max(crime$Reported_Date)
```

```
## [1] "2023-01-01 23:56:46 UTC"
```

```
hist(crime$Reported_Date, breaks="month")
```



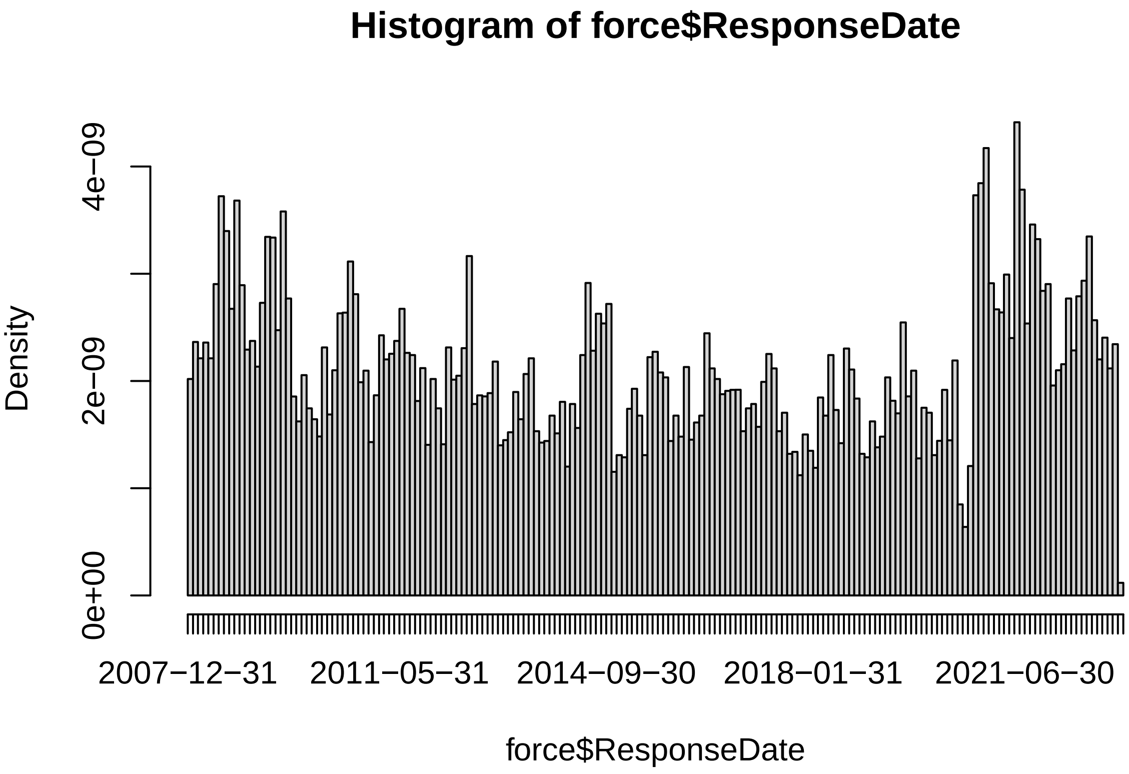
```
min(force$ResponseDate)
```

```
## [1] "2008-01-01 01:15:31 UTC"
```

```
max(force$ResponseDate)
```

```
## [1] "2023-01-01 02:29:28 UTC"
```

```
hist(force$ResponseDate, breaks="month")
```



Understanding setdiff

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

```
a <- c(1, 2, 3, 4, 5)
b <- c(2, 3, 4, 6)
c(setdiff(a, b), intersect(a, b))
```

```
## [1] 1 5 2 3 4
```

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 use of force data

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

```
# crime %>%
#     filter(Case_Number %in% setdiff(crime$Case_Number, force$CaseNumber)) %>%
#     ggplot() +
#     geom_point(aes(x=X, y=Y))
```

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

```
## [1] 0.01092098
```

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

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

```
## [1] 0.06173242
```

Comparing crime data and incidents data

```
incidents_caseNumber <- c(
    incidents_2019$caseNumber,
    incidents_2020$caseNumber,
    incidents_2021$caseNumber,
    incidents_2022$caseNumber
    )
length(intersect(incidents_caseNumber, crime$Case_NumberAlt)) / length(incidents_caseNumber)
```

```
## [1] 0.2743099
```

```
length(intersect(crime$Case_NumberAlt, incidents_caseNumber)) / length(crime$Case_NumberAlt)
```

```
## [1] 0.1276245
```