NYPD Shooting Incidents

Osher Jacob

2022-09-29

NYPD Shooting Incidents

Data description

List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included.

Preparing our data

```r
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

incidents_raw <- read.csv(url)
```

```r
# Removing unnecessary fields, and converting times & dates
incidents <- incidents_raw %>%
  select(
    -c(Latitude, Longitude, X_COORD_CD, Y_COORD_CD,
       Lon_Lat,JURISDICTION_CODE, STATISTICAL_MURDER_FLAG)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE), OCCUR_TIME = hms(OCCUR_TIME))

# Add time hour column
incidents$hour <- hour(incidents$OCCUR_TIME)

# Handle missing data

incidents$PERP_SEX[incidents$PERP_SEX == ""] <- "U"
incidents$PERP_RACE[incidents$PERP_RACE == ""] <- "UNKNOWN"
incidents$PERP_AGE_GROUP[incidents$PERP_AGE_GROUP == ""] <- "UNKNOWN"
incidents$PERP_AGE_GROUP[incidents$PERP_AGE_GROUP %in% c(940,224,1020)] <- "UNKNOWN"
incidents$LOCATION_DESC[incidents$LOCATION_DESC == ""] <- "UNKNOWN"


# Converting categorical variables to factors
incidents$LOCATION_DESC <- as.factor(incidents$LOCATION_DESC)
incidents$BORO <- as.factor(incidents$BORO)
incidents$PERP_RACE <- as.factor(incidents$PERP_RACE)
incidents$PERP_SEX <- as.factor(incidents$PERP_SEX)
```

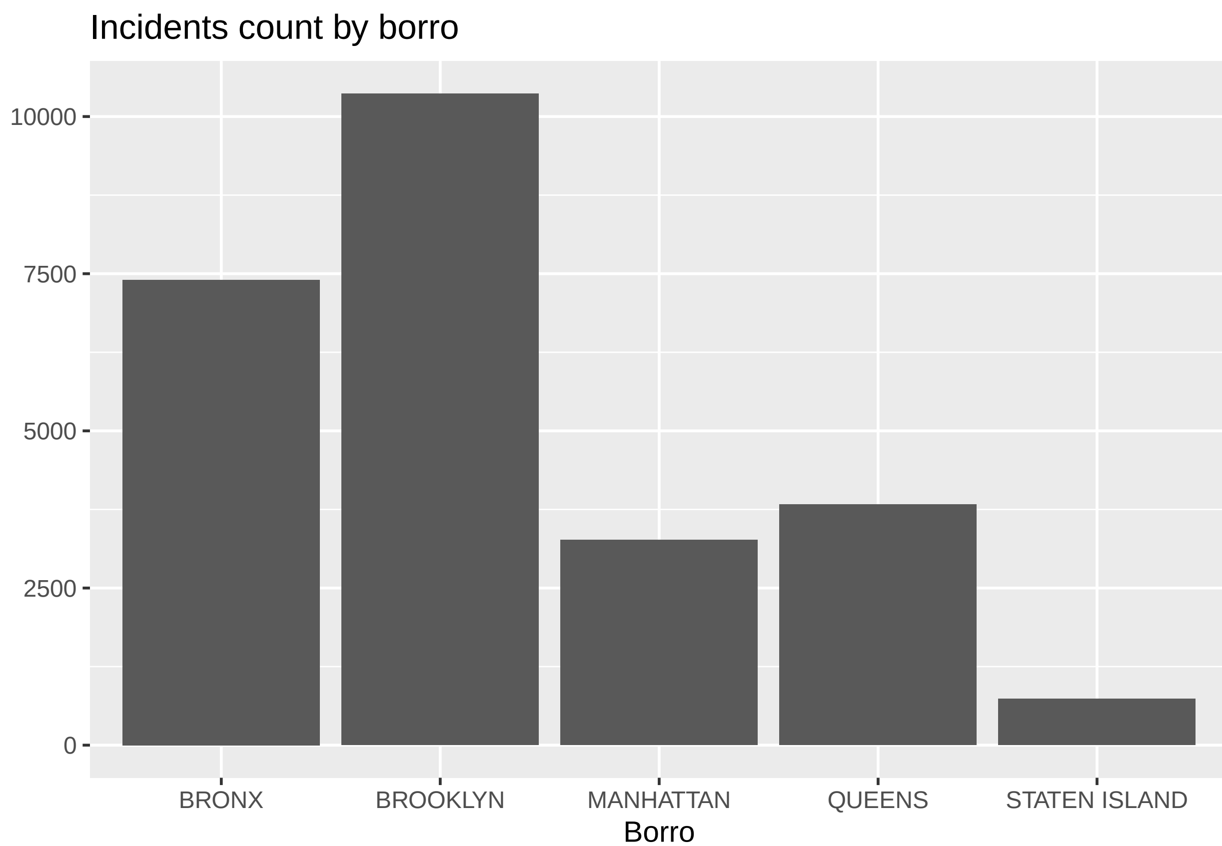
```r
incidents$PERP_AGE_GROUP <- as.factor(incidents$PERP_AGE_GROUP)
incidents$PRECINCT <- as.factor(incidents$PRECINCT)
incidents$VIC_RACE <- as.factor(incidents$VIC_RACE)
incidents$VIC_SEX <- as.factor(incidents$VIC_SEX)
incidents$VIC_AGE_GROUP <- as.factor(incidents$VIC_AGE_GROUP)
```

```r
summary(incidents)
```

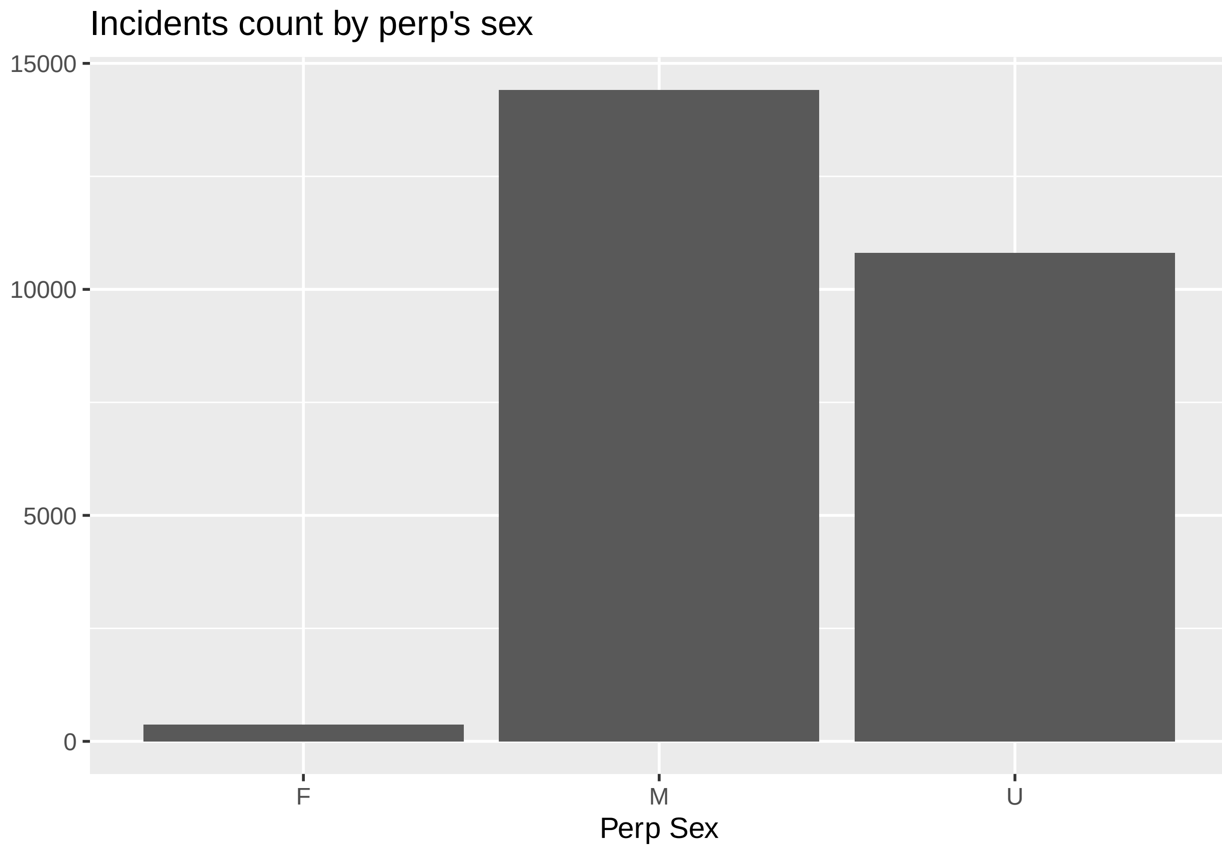
Visualization and analysis

```r
# MAke some plots

as.data.frame(table(incidents$BORO)) %>%
  ggplot(aes(x=Var1, y=Freq)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "bottom") +
  labs(title = "Incidents count by borro", y=NULL, x='Borro')
```




```r
as.data.frame(table(incidents$PERP_SEX)) %>%
  ggplot(aes(x=Var1, y=Freq)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "bottom") +
  labs(title = "Incidents count by perp's sex", y=NULL, x='Perp Sex')
```



By looking at the perpetrator sex bar chart above, it seems that there's an enormous difference between males and females as the reported perpetrator sex. This leads to the question: Are females less involved in shooting incidents at all (in NYC)? Or are they less likely to be the perpetrator of such an incident?

To examine further we can make a similar chart of the victim's sex, and see if we get the same ratio of males to females as before.

The perpetrator male to female ratio is:

```
length(incidents$PERP_SEX[incidents$PERP_SEX == "M"]) /
length(incidents$PERP_SEX[incidents$PERP_SEX == "F"])
```

```
## [1] 38.85714
```

Which means for every reported female shooter we have around 39 male shooters.

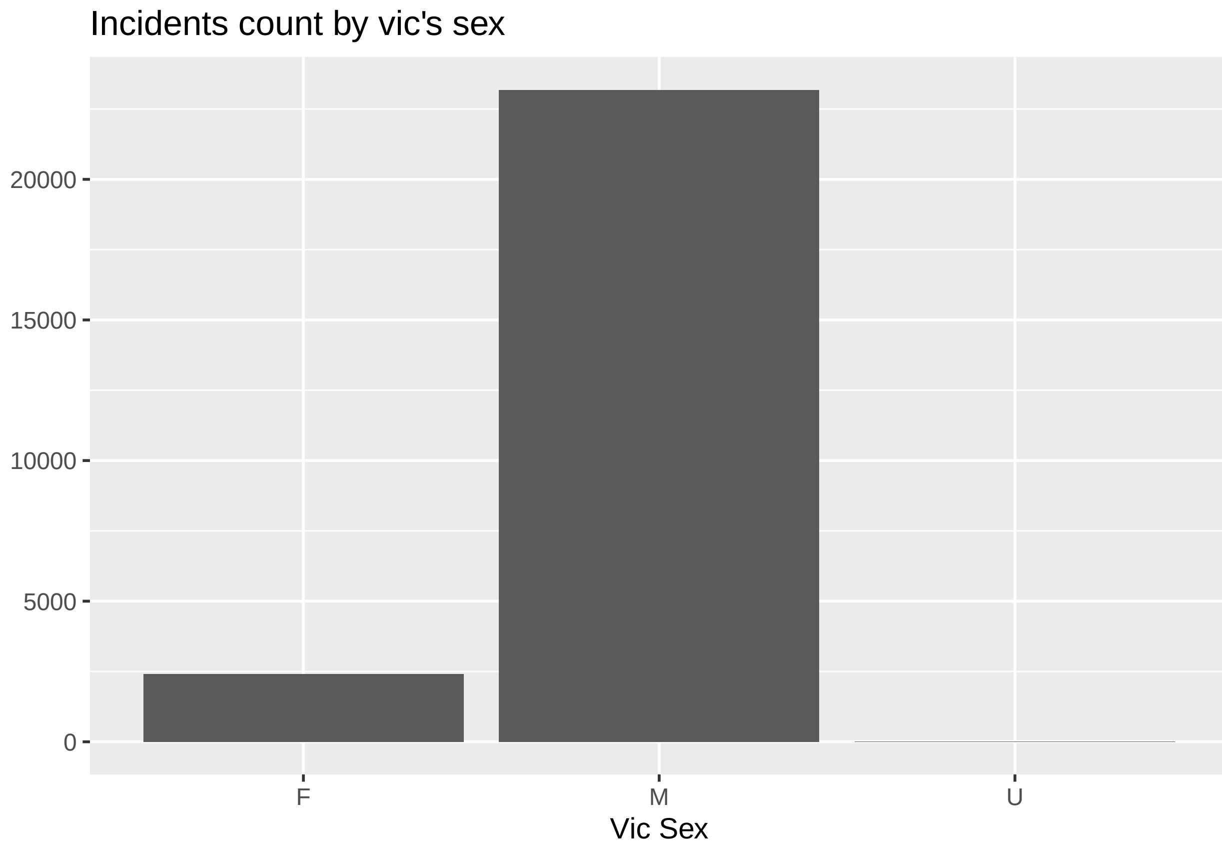
The male perpetrator to victim ratio:

```
length(incidents$PERP_SEX[incidents$PERP_SEX == "M"]) /
length(incidents$PERP_SEX[incidents$VIC_SEX == "M"])
```

```
## [1] 0.6218618
```

```
# MAke another plot

as.data.frame(table(incidents$VIC_SEX)) %>%
  ggplot(aes(x=Var1, y=Freq)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "bottom") +
  labs(title = "Incidents count by vic's sex", y=NULL, x='Vic Sex')
```



The victim male to female ratio is:

```
length(incidents$VIC_SEX[incidents$VIC_SEX == "M"]) /
length(incidents$VIC_SEX[incidents$VIC_SEX == "F"])
```

```
## [1] 9.647108
```

Which means for every reported female victim, we have around 10 male victims which is lower than our perpetrator ratio.

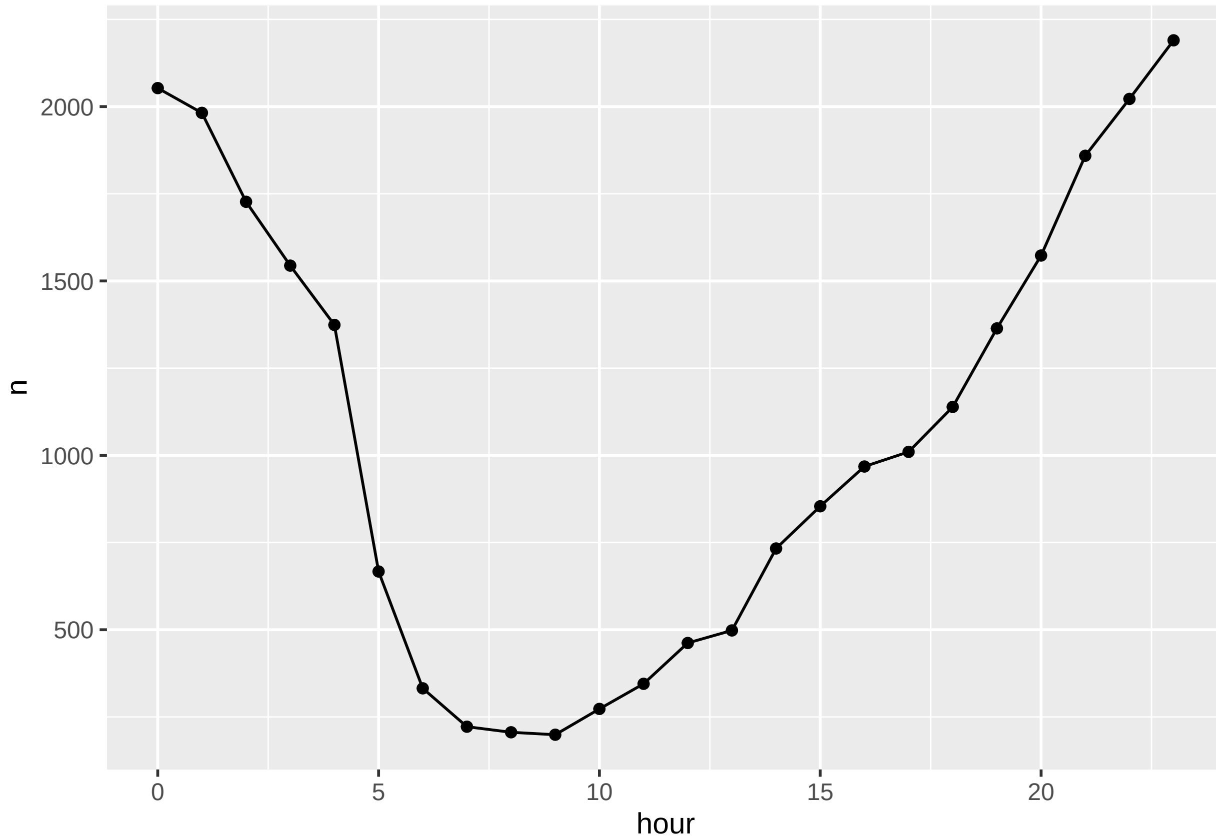
The female perpetrator to victim ratio:

```
length(incidents$PERP_SEX[incidents$PERP_SEX == "F"]) /
length(incidents$PERP_SEX[incidents$VIC_SEX == "F"])
```

```
## [1] 0.1543903
```

Let us examine a different topic, which is the occurrence time of the incidents.

```
# Plot common hours of incidents
incidents_per_hour = incidents %>% count(hour)
incidents_per_hour %>% ggplot(aes(x=hour, y=n)) + geom_point() + geom_line()
```

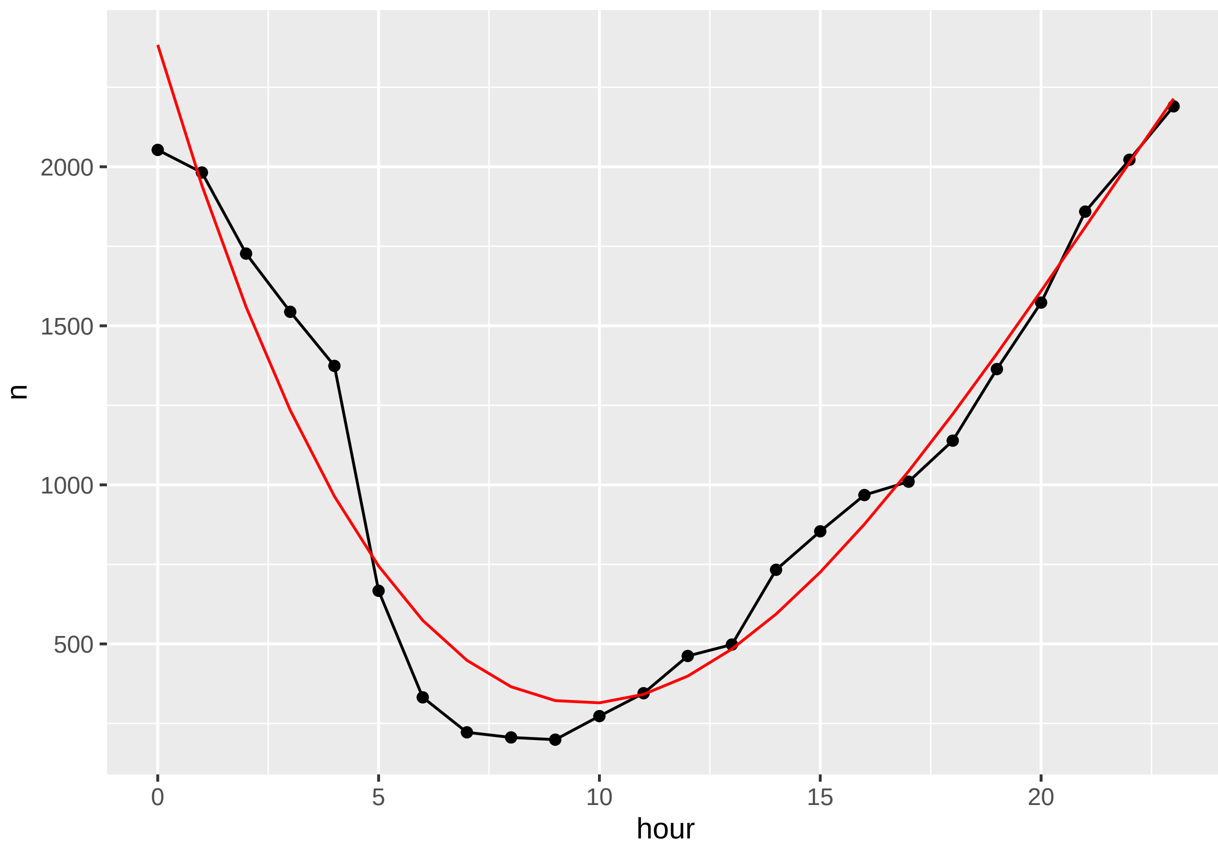


In this chart we can clearly see that most incidents occur at night. We can try and fit a model through this data, in order to determine the incident count at the different hours of the day.

```
model <- lm( incidents_per_hour$n ~ poly(incidents_per_hour$hour,3) )

incidents_per_hour_w_pred = incidents %>% count(hour)
incidents_per_hour_w_pred$pred <- predict(model, x=incidents_per_hour_w_pred$hour)

incidents_per_hour_w_pred %>% ggplot(aes(x=hour, y=n)) +
  geom_point() +
  geom_line() +
  geom_line(aes(x=hour, y=pred), color="red")
```



The prediction of the model (red line), seems to fit the curve quite well.

Conclusions

Regarding our incident time analysis, it appear that the safest hours are between 6-11 am, and the most violent are between 10pm to 2am.

Regarding sex of the people involved in shooting incidents. In general, females are less likely to be involved in our reported shooting incidents, and when they are, they're usually on the victim side rather than the perpetrator.

Some thing to notice about this data and be cautious about, is that we have a big number of unknown sex in our perpetrators. This raises further questions about the validity of these numbers.

Possible sources of bias

These examinations and results lean heavily on how the data was collected and reported. As mentioned before, the sex of the perpetrator is mostly unknown, this could create a huge bias in the data set. The data is only incidents that occurred in NYC, making general conclusions about other places from this data is also subject to heavy bias.

Another disclaimer about creator of this project. I am a male, and my experience of NYC as a tourist was entertaining and non-violent. This means I could've misinterpreted the results one way or another due to personal bias.

```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
## Running under: Windows 10 x64 (build 14393)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.8.0 forcats_0.5.2   stringr_1.4.0   dplyr_1.0.10
##  [5] purrr_0.3.4     readr_2.1.2     tidyr_1.2.0     tibble_3.1.7
##  [9] ggplot2_3.3.6   tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.2    xfun_0.31           haven_2.5.1
##  [4] gargle_1.2.0        colorspace_2.0-3    vctrs_0.4.1
##  [7] generics_0.1.3      htmltools_0.5.2     yaml_2.3.5
## [10] utf8_1.2.2          rlang_1.0.3         pillar_1.7.0
## [13] withr_2.5.0         glue_1.6.2          DBI_1.1.3
## [16] dbplyr_2.2.1        readxl_1.4.1        modelr_0.1.9
## [19] lifecycle_1.0.1     munsell_0.5.0       gtable_0.3.1
## [22] cellranger_1.1.0    rvest_1.0.3         evaluate_0.15
## [25] labeling_0.4.2      knitr_1.39          tzdb_0.3.0
## [28] fastmap_1.1.0       fansi_1.0.3         highr_0.9
## [31] broom_1.0.1         backports_1.4.1     scales_1.2.1
## [34] googlesheets4_1.0.1 jsonlite_1.8.0      farver_2.1.1
## [37] fs_1.5.2            hms_1.1.2           digest_0.6.29
## [40] stringi_1.7.6       grid_4.2.1          cli_3.3.0
## [43] tools_4.2.1         magrittr_2.0.3      crayon_1.5.1
## [46] pkgconfig_2.0.3     ellipsis_0.3.2      xml2_1.3.3
## [49] reprex_2.0.2        googledrive_2.0.0   assertthat_0.2.1
## [52] rmarkdown_2.16      httr_1.4.3          rstudioapi_0.13
## [55] R6_2.5.1            compiler_4.2.1
```