DTSA 5301, week3 - NYPD project

Serg Prokhorov

2023-06-20

NYPD dataset simple analysis

Project goal

This project summarises knowledge acquired during week 3 of **Data Science as a Field** course from CU Boulder. As we were mostly focused on introducing the basic functionality of R Markup and R Studio environment in the course so far, the following document serves mostly to demonstrate basic data analysis approaches, without deep reliance on the data meaning. I assume we'll address this topic on later stages of our education.

Data source

The source files for the project are from official U.S. Government's Open Data repository https://catalog.data.gov/dataset, specifically the dataset titled **NYPD Shooting Incident Data (Historic)**.

Data file address is https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD and it can be queried on-line later for report results reproducibility.

Load data from URLs

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

nypd_data <- read_csv(nypd_url, show_col_types = FALSE)
```

Summary of source NYPD data

Below is a summary of loaded dataset structure:

```
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
```

```
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Clean up of NYPD data

To clean up source data we perform following transformations:

- convert date field (OCCUR_DATE) format from text to date
- remove unused in further analysis columns (date, time, geo-data, misc. attributes)

```
nypd_data_clean <- nypd_data %>%
  # date type conversion
  mutate(Date = mdy(OCCUR_DATE)) %>%
```

```
  # remove columns
  select(-c(
    # date - time, already extracted the date
    OCCUR_DATE, OCCUR_TIME,
    # geo part - not needed
    X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat,
    # not needed now (probably)
    LOC_CLASSFCTN_DESC, INCIDENT_KEY
    )
  )
```

Summary of data after cleaning up

```
summary(nypd_data_clean)
```

```
##      BORO           LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE
##  Length:27312       Length:27312       Min.   :  1.00   Min.   :0.0000
##  Class :character   Class :character   1st Qu.: 44.00   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median : 68.00   Median :0.0000
##                                        Mean   : 65.64   Mean   :0.3269
##                                        3rd Qu.: 81.00   3rd Qu.:0.0000
##                                        Max.   :123.00   Max.   :2.0000
##                                                         NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE             Date
##  Length:27312       Min.   :2006-01-01
##  Class :character   1st Qu.:2009-07-18
##  Mode  :character   Median :2013-04-29
##                     Mean   :2014-01-06
##                     3rd Qu.:2018-10-15
##                     Max.   :2022-12-31
##
```

If some data is missing on later stage of project, I will come back and correct the transformation procedures, probably adding missing data from additional datasets, how it was shown in lecture with Population data.

Data visualization and analysis

To summarize and enrich data following steps are performed:

- Aggregate accident cases by territory and date into **nypd_by_terr**
- Summarize all territories by date into **nypd_by_date**

The **main distinction of this analysis** from shown in the lectures is that in addition to summing or finding extremes (min/max) during aggregation, now we perform counting of rows merged during aggregation using the *n()* function.

```
nypd_by_terr <- nypd_data_clean %>%
  group_by( BORO, Date) %>%
  summarise(cases = n() ) %>% # n() to count rows in group
  ungroup()
```

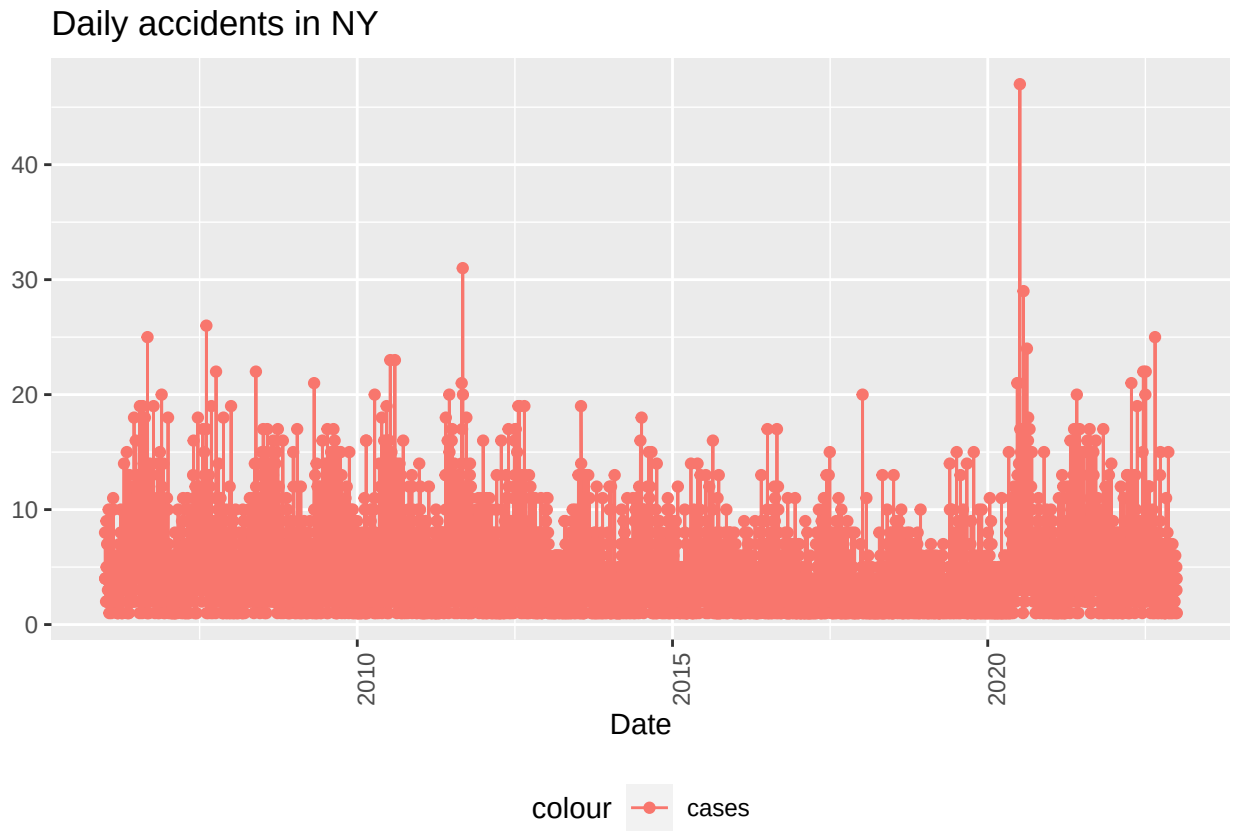
```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

```
nypd_by_date <- nypd_by_terr %>%
  group_by(Date) %>%
  summarize(cases = sum(cases))
```

Simple visualization is provided below to get a first glance on the nature of the data aggregation results:

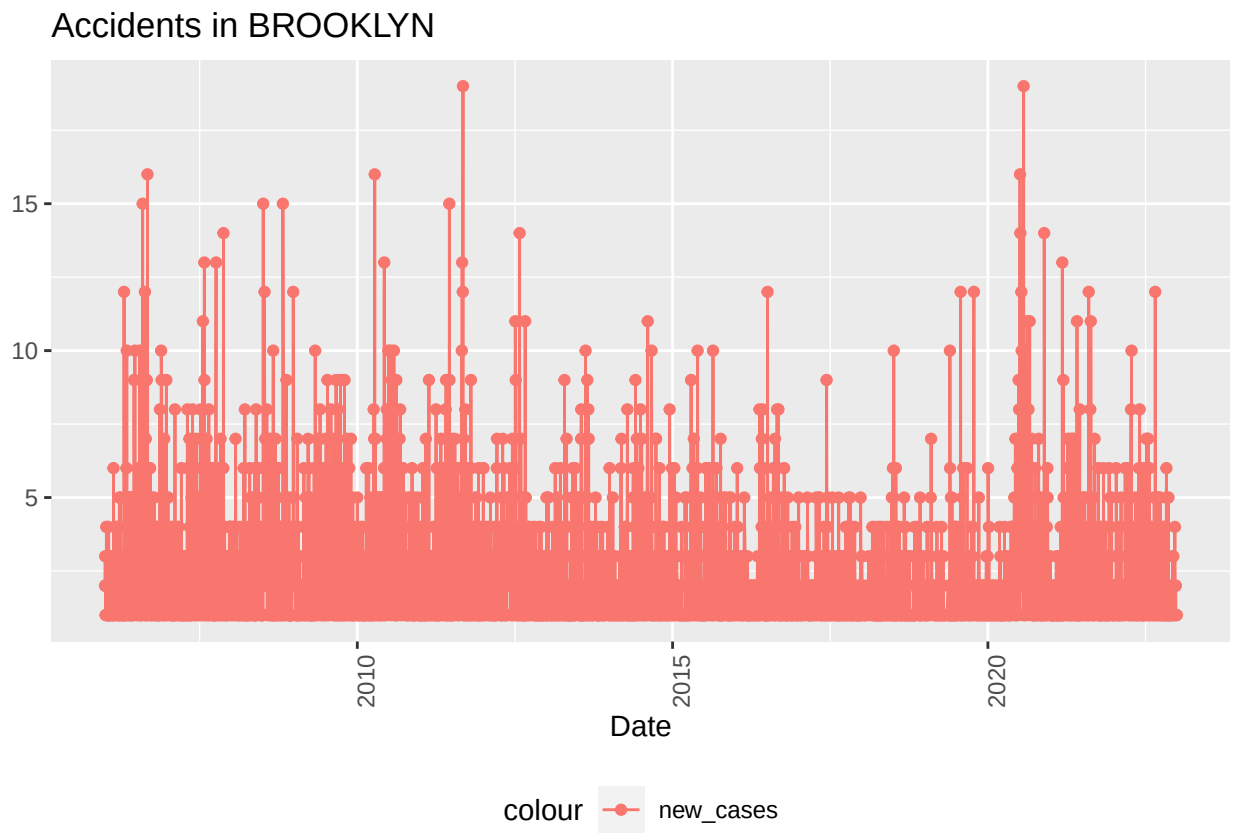
```
nypd_by_date %>%
  filter(cases > 0) %>%
  ggplot(aes(x = Date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +

  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Daily accidents in NY", y = NULL)
```




```
terr <- "BROOKLYN"

nypd_by_terr %>%
  filter(cases > 0) %>%
  filter(BORO == terr) %>%
  ggplot(aes(x = Date, y = cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color = "new_cases")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("Accidents in ", terr), y = NULL)
```



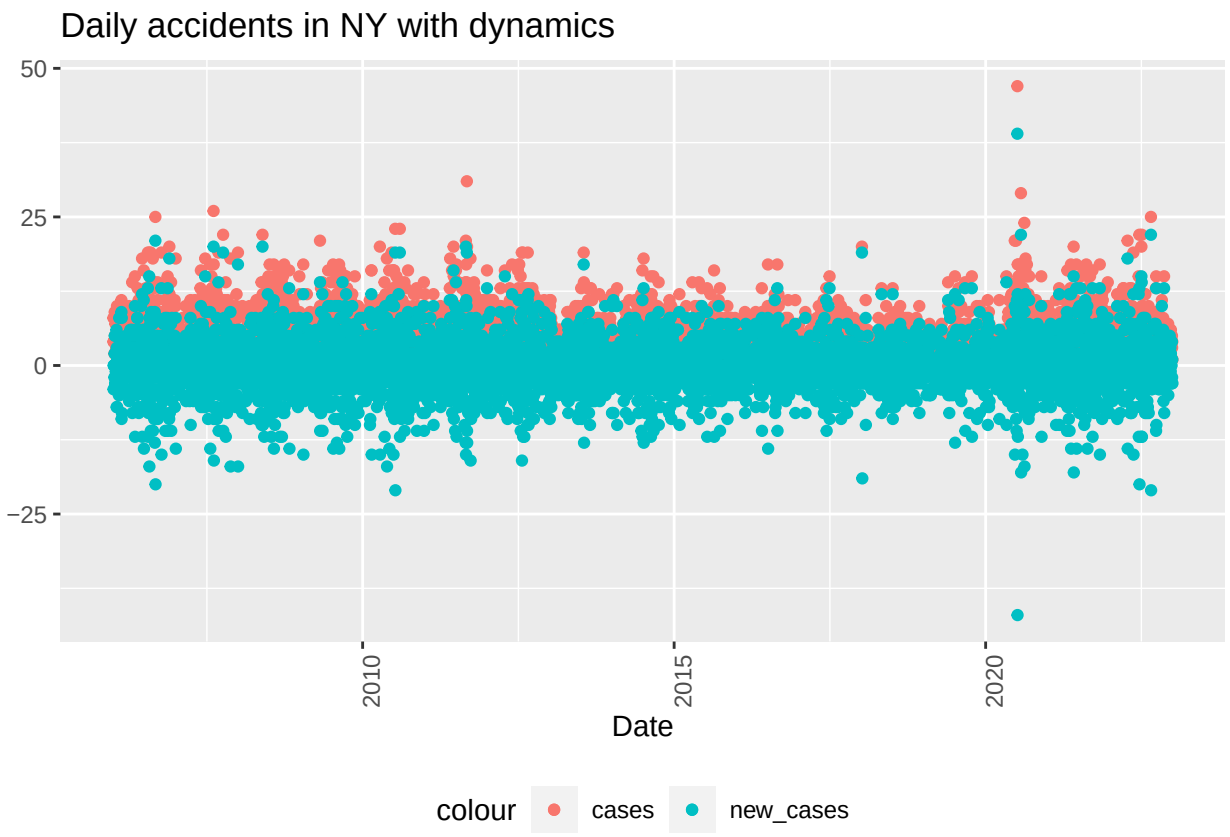
Further analysis

For further analysis we enrich source data with calculation of daily dynamics (difference with previous day) for number of accidents, and visualizing the result:

```
nypd_by_date <- nypd_by_date %>%
  mutate(NewCases = cases - lag(cases))

nypd_by_date %>%
  ggplot(aes(x = Date, y = cases)) +
  geom_point(aes(color = "cases")) +
  geom_point(aes(y = NewCases, color = "new_cases")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Daily accidents in NY with dynamics", y = NULL)
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```



First conclusions

Looking at the new cases/accidents we can see repeatable trends of growth and decline in cases dynamics, which probably can be further analysed to either identify source data discrepancies, or by adding additional factors into analysts, try to identify additional dependencies.

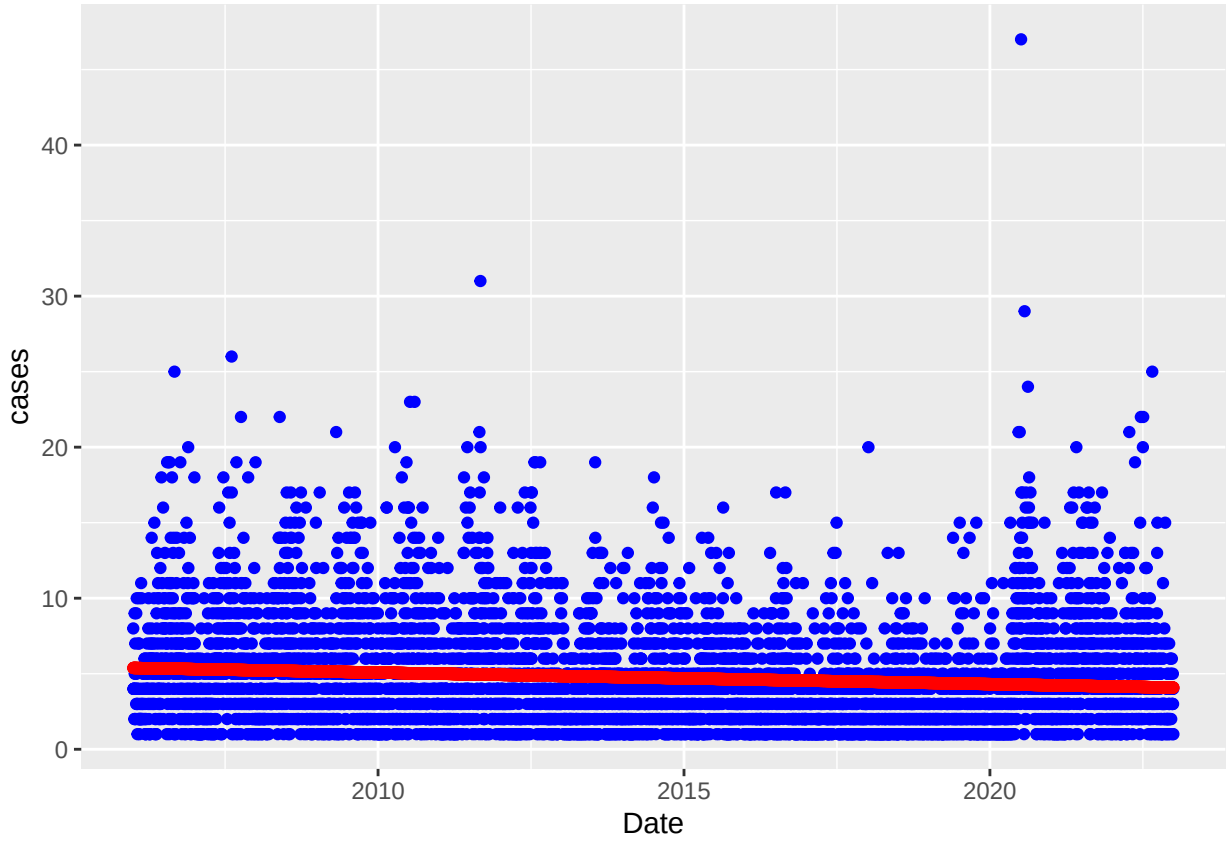
Predictive modeling of data

We'll build a model to predict cases count and visualize both predicted and actual values on same graph

```
mod <- lm(cases ~ Date, data = nypd_by_date)

nypd_by_date_w_pred <- nypd_by_date %>%
  mutate(pred = predict(mod))

nypd_by_date_w_pred %>%
  ggplot(  ) +
  geom_point(aes(x = Date, y = cases), color = "blue") +
  geom_point(aes(x = Date, y = pred), color = "red")
```



Looking on these two graphs, it's clear that the model type used provides over-simplified representation of data trends in source data, although there's definitely a correlation present. I assume further courses will introduce us to more complicated modeling techniques, allowing to get more correct predictions.

Conclusions and Bias Identification

Conclusion to the project report

Working on the NYPD data set, because of significant amount of data attributes collected with each incident, demonstrated several possibilities to analyse data by grouping various attributes. Also, the fact that we removed most of existing attributes, for the sake of simple demonstration of data processing concepts taught in this class, hints that there are many opportunities for additional analysis, would the task at hand be more related to real world needs - for example, use of demographics or spatial data.

Possible sources of bias

Bias can appear from personal beliefs of the data scientist performing the analysis, also the way the source data was gathered, and how the report was designed, its goals and requested analysis criteria from the customer. All this can significantly influence the outcome. Usually bias comes from deep beliefs, based on ancient survival mechanisms. They usually influence someones decisions on unconscious level, and additional steps needs to be taken to identify and prevent bias.

Possible personal bias in the analysis

I assume my specific gender, race, previous knowledge of some city districts rumors (safety, wealth) and similar beliefs, could have impacted the way I approached this project.

Personal bias mitigation steps taken

Knowing that some topics could be biased I took additional steps to ensure that my analysis treats them fairly and universally. For example, when doing aggregation by city districts, I ensured that all of them were analysed equally, without adding any additional weights or parameters, not relevant to the study performed.

Appendix A - session info

The report was generated using the following software/libraries:

```
sessionInfo()
```

```
## R version 4.3.0 (2023-04-21)
## Platform: x86_64-apple-darwin20 (64-bit)
## Running under: macOS Monterey 12.6.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: Asia/Dubai
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.2 forcats_1.0.0   stringr_1.5.0   dplyr_1.1.2
##  [5] purrr_1.0.1     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1
##  [9] ggplot2_3.4.2   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5        gtable_0.3.3     highr_0.10       crayon_1.5.2
##  [5] compiler_4.3.0   tidyselect_1.2.0 parallel_4.3.0   scales_1.2.1
##  [9] yaml_2.3.7       fastmap_1.1.1    R6_2.5.1         labeling_0.4.2
## [13] generics_0.1.3   curl_5.0.1       knitr_1.43       munsell_0.5.0
## [17] pillar_1.9.0     tzdb_0.4.0       rlang_1.1.1      utf8_1.2.3
## [21] stringi_1.7.12   xfun_0.39        bit64_4.0.5      timechange_0.2.0
## [25] cli_3.6.1        withr_2.5.0      magrittr_2.0.3   digest_0.6.31
## [29] grid_4.3.0       vroom_1.6.3      rstudioapi_0.14  hms_1.1.3
## [33] lifecycle_1.0.3  vctrs_0.6.2      evaluate_0.21    glue_1.6.2
## [37] farver_2.1.1     fansi_1.0.4      colorspace_2.1-0 rmarkdown_2.22
## [41] tools_4.3.0      pkgconfig_2.0.3  htmltools_0.5.5
```