NYPD Shootings

W. Pericles

2025-07-27

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

NYPD Shooting Dataset

The NYPD Shooting data set is comprised of every shooting incident in NYC going back as far as 2006 up to 2024. Each record represents a shooting incident and contains information related to date and time, information about the shooting event, victim and suspect demographics, and what borough the incident took place.

```
csv_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_csv <- read_csv(csv_url)
```

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): X_COORD_CD, Y_COORD_CD
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Summary of Data and Cleaning

```
summary(nypd_csv)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:29744       Length:29744       Length:29744
##  1st Qu.: 67321140   Class :character   Class1:hms         Class :character
##  Median :109291972   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :133850951                      Mode  :numeric
##  3rd Qu.:214741917
##  Max.   :299462478
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:29744       Min.   :  1.00   Min.   :0.0000    Length:29744
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.00   Median :0.0000    Mode  :character
##                     Mean   : 65.23   Mean   :0.3181
##                     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:29744       Mode :logical           Length:29744
##  Class :character   FALSE:23979             Class :character
##  Mode  :character   TRUE :5765              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:29744       Length:29744       Length:29744       Length:29744
##  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:29744       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000094   1st Qu.:183042   1st Qu.:40.67
##  Mode  :character   Median :1007826   Median :195506   Median :40.70
##                     Mean   :1009442   Mean   :208722   Mean   :40.74
##                     3rd Qu.:1016739   3rd Qu.:239980   3rd Qu.:40.83
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :97
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:29744
##  1st Qu.:-73.94   Class :character
##  Median :-73.91   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :97
```

There are many columns of data that is not necessary for the visualization and analysis I want to perform. In cleaning the data I removed the Incident Key, all data in between Precinct and Location Description,

and from 'X_COORD' to 'Lon_Lat'. With the remaining columns there were still plenty of NAs and nulls. Many of the missing data came from the columns related to the perpetrator. My assumption is that the perpetrators that carried the act that led to missing data were never caught. For now those values will be replaced with "Unknown".

```
nypd_csv <- nypd_csv %>% select(-INCIDENT_KEY, -c(PRECINCT:LOCATION_DESC),
                         -c(X_COORD_CD:Lon_Lat)) %>%
                         mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
                         rename(Murder = `STATISTICAL_MURDER_FLAG`) %>%
         replace_na(list(LOC_OF_OCCUR_DESC = "Unknown", PERP_AGE_GROUP = "Unknown",
         PERP_SEX = "Unknown",PERP_RACE = "Unknown"))
nypd_csv$PERP_AGE_GROUP[nypd_csv$PERP_AGE_GROUP == '(null)'] <- "Unknown"
nypd_csv$PERP_SEX[nypd_csv$PERP_SEX == '(null)'] <- "Unknown"
nypd_csv$PERP_RACE[nypd_csv$PERP_RACE == '(null)'] <- "Unknown"

summary(nypd_csv)
```

```
##    OCCUR_DATE          OCCUR_TIME           BORO           LOC_OF_OCCUR_DESC
##  Min.   :2006-01-01   Length:29744      Length:29744       Length:29744
##  1st Qu.:2009-10-29   Class1:hms        Class :character   Class :character
##  Median :2014-03-25   Class2:difftime   Mode  :character   Mode  :character
##  Mean   :2014-10-31   Mode  :numeric
##  3rd Qu.:2020-06-29
##  Max.   :2024-12-31
##    Murder         PERP_AGE_GROUP       PERP_SEX          PERP_RACE
##  Mode :logical   Length:29744       Length:29744       Length:29744
##  FALSE:23979     Class :character   Class :character   Class :character
##  TRUE :5765      Mode  :character   Mode  :character   Mode  :character
##
##
##
##  VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:29744       Length:29744       Length:29744
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Visualization and Analysis

2020 and Incidents During Holidays

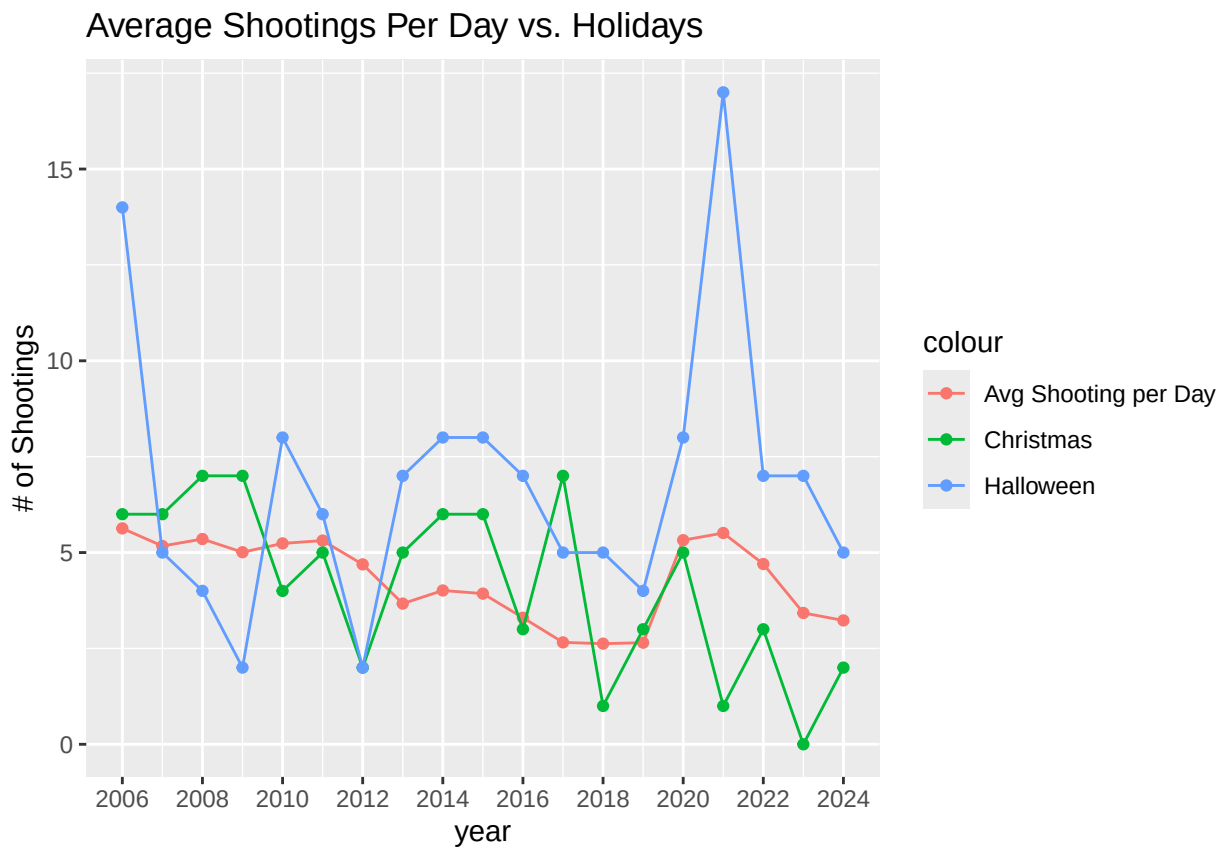
Naturally, I wanted to know what was the average number of shootings per day per year. I wanted to see if the average would rise over time, decline, or stay about the same level. I was pleased to see the that from 2006 to 2019 that shooting incidents had a decline. But in 2020, the shootings incidents climbed back up to the levels from when the data was first being tracked. I suspected that the 2020 COVID-19 pandemic and social unrest, such as the killing of George Floyd, was the cause of the rise in shootings. 2020 did see a record number of gun purchases. Thankfully the rate of shootings have since declined.

After viewing the data for average number of shootings per day, I wanted to know if holidays significantly change the number of shootings in the day. From the available data, I counted the number of shootings that

occurred in each Christmas and Halloween. Roughly, about half of the years showed that there was more shootings than the average for a given year on Christmas, with 2017 having a well above averge number of shootings that day. I expected for Christmas to be exclusively a below average shooting day due to the holiday's generally positive spirit. 2023 did have the Christmas miracle of have no reported shootings!

However, Halloween has far more "above the daily average" shootings. 2021 had 17 shooting incidents on Halloween night, while the daily average was 5.5 shootings a day. Halloween involves the outdoot activites like trick-or-treating and pumpkin patches. With the increase of interactions between people you can expect the chances of shootings go up.

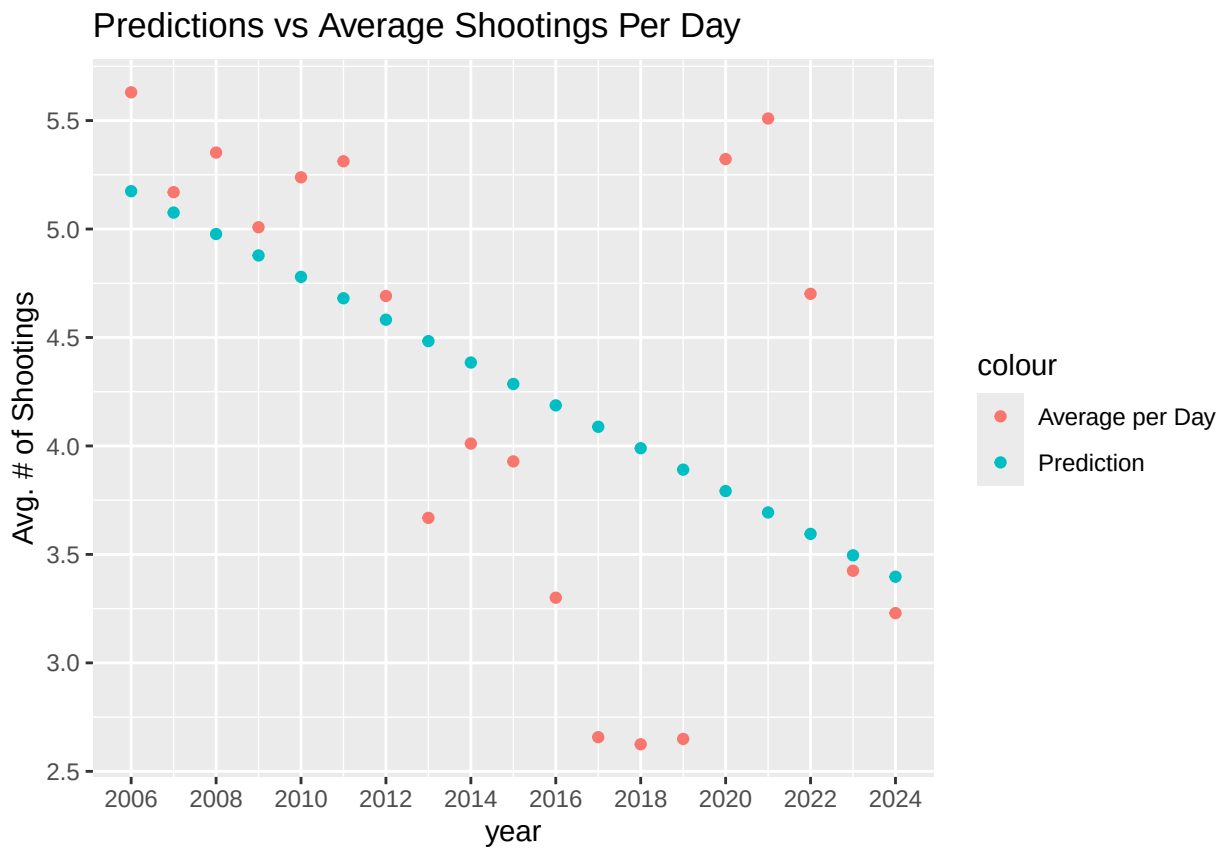
```
shooting_per_year <- nypd_csv %>% group_by(year = year(OCCUR_DATE)) %>%
          select(OCCUR_DATE, year) %>%
          count(year, name = "Incidents") %>%
          mutate(avg = ifelse(year %%4 == 0, Incidents/366,Incidents/365))

# Number of shootings on Christmas
christmas_incidents <- nypd_csv %>% filter(format(OCCUR_DATE, "%m") == "12"
                & format(OCCUR_DATE, "%d") == "25") %>%
                group_by(year = year(OCCUR_DATE)) %>%
                select(OCCUR_DATE, year) %>%
                count(year, name = "Incidents") %>%
                tibble() %>% add_row(year = 2023, Incidents = 0)
christmas_incidents <- christmas_incidents %>% sort_by(christmas_incidents,
                      christmas_incidents$year)

# Number of shootings on Halloween
halloween <- nypd_csv %>% filter(format(OCCUR_DATE, "%m") == "10" &
                          format(OCCUR_DATE, "%d") == "31") %>%
                          group_by(year = year(OCCUR_DATE)) %>%
                          select(OCCUR_DATE, year) %>%
                          count(year, name = "Incidents")

# plotting data
shooting_per_year %>% ggplot(aes(x = year, y = avg)) +
      geom_line(aes(color = "Avg Shooting per Day")) +
      geom_point(aes(color = "Avg Shooting per Day")) +
      geom_line(aes(y = christmas_incidents$Incidents, color = "Christmas")) +
      geom_point(aes(y = christmas_incidents$Incidents, color = "Christmas")) +
      geom_line(aes(y = halloween$Incidents, color = "Halloween")) +
      geom_point(aes(y = halloween$Incidents, color = "Halloween"))+
      scale_x_continuous(breaks=seq(2006,2024,by=2)) +
      labs(title = "Average Shootings Per Day vs. Holidays", y = "# of Shootings",
           theme(plot.title = element_text(hjust = 0.5)))
```



Modeling

From 2006 to 2019, the average shootings per day made a downward trend, showing great improvement. But with the 2020 pandemic and social unrest, the average shooting per day skyrocketed. The prediction model below illustrates the average shooting per day would be about 3.8 a day, but the actual data shows that the average was 5.3. Years 2021 and 2022 were also well above the prediction model, but 2023 and 2024 show that shooting incidents are beginning to cool off again as we get further from the pandemic years.

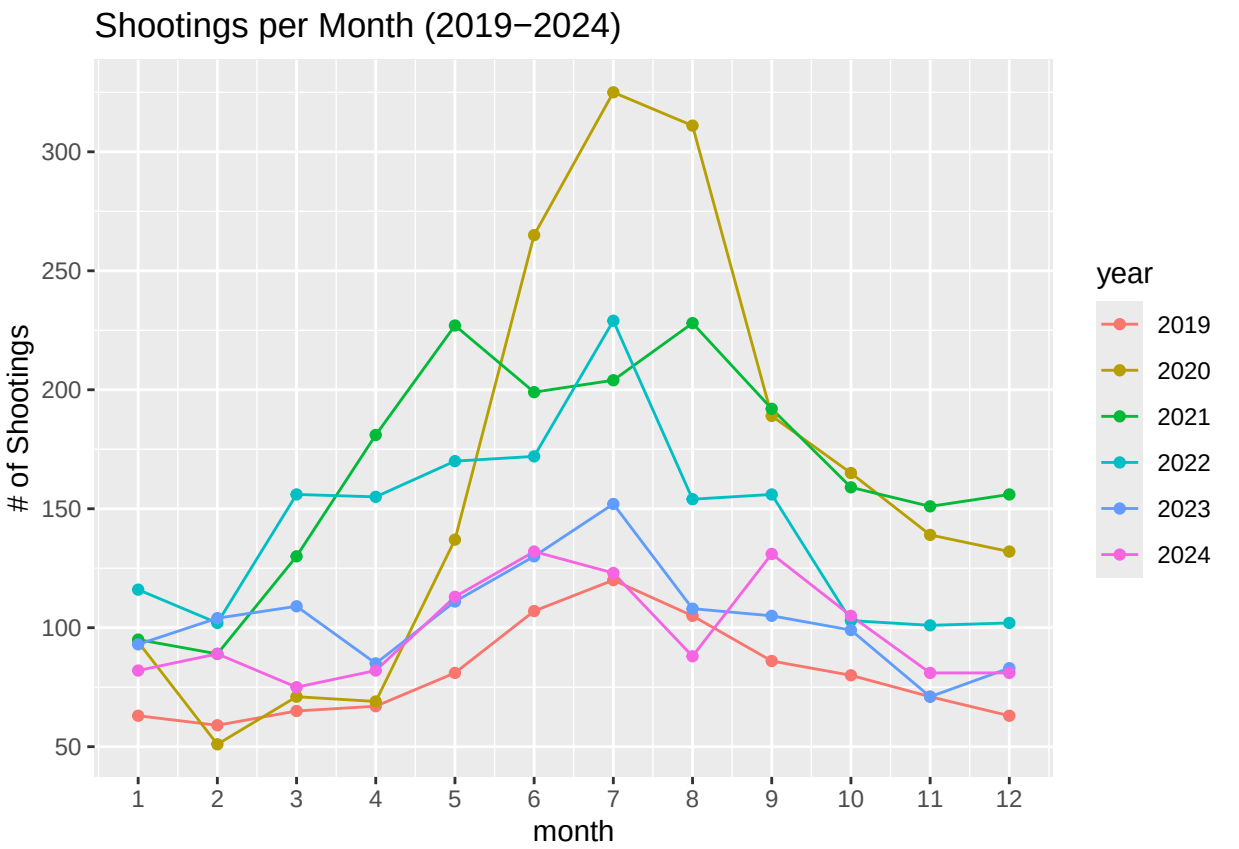
```
model <- shooting_per_year %>% lm(formula = avg ~ year)
pred <- predict(model, tibble(year = c(2006:2024)))
shooting_per_year %>% ggplot(mapping = aes(x = year, y = pred)) +
  geom_point(mapping = aes(color = "Prediction")) +
  geom_point(mapping = aes(y = avg, color = "Average per Day")) +
  scale_y_continuous(breaks=seq(2,6,by=0.5)) +
  scale_x_continuous(breaks=seq(2006,2024,by=2)) +
  labs(title = "Predictions vs Average Shootings Per Day",
       y = "Avg. # of Shootings")
```



Incidents Month Over Month

When I started to look at the data, I started to as myself, "are the number of shootings distributed evenly throughout the year, or do shootings occur more often in certain times of the year?" So I organized the data to count the number of shootings in a given month in each respective year. Below is a plot of shootings from 2019-2023. I immediately noticed that there are more shootings in the summer months, but a decline as the year goes into the cooler months.

```
month_by_month <- nypd_csv %>% group_by(year = year(OCCUR_DATE),
                  month = month(OCCUR_DATE)) %>%
                  select(year, month) %>%
                  count(year, month)

month_by_month %>% filter(year>=2019) %>%
                  mutate(year = as.character(year)) %>%
                  ggplot(mapping = aes(x = month, y = n)) +
                  geom_line(mapping = aes(color = year)) +
                  geom_point(mapping = aes(color = year)) +
                  scale_x_continuous(breaks=seq(1,12,by=1)) +
                  scale_y_continuous(breaks=seq(0,400,by=50)) +
  labs(title = "Shootings per Month (2019-2024)", y = "# of Shootings")
```



Conclusion

The rate of shootings that occur in NYC can change over the course of a year or be affected by world events. Not only did the COVID-19 pandemic cause a health crisis, it caused a crime crisis as well. The data also suggests that holidays, especially Halloween, can have higher number of incidents than the average, and the temperature can affect the numbers too.

Bias

I did go into this research with the biased opinion that Christmas would always have a low number of shooting incidents due to its generally positive presence. I was disappointed to see that about half of the years from the data have the number of Christmas shootings higher than the daily average. To mitigate the disappointment, I added the Halloween data to the plot, assuming that it would have a higher number of shootings. The high Halloween numbers can lead people to not think so poorly about the Christmas numbers. This is an example of poor ethics.

Session Info

```
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31)
## Platform: x86_64-apple-darwin20
## Running under: macOS Ventura 13.7.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-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: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.4     readr_2.1.5     tidyr_1.3.1     tibble_3.3.0
##  [9] ggplot2_3.5.2   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.6.0          gtable_0.3.6       crayon_1.5.3       compiler_4.4.2
##  [5] tidyselect_1.2.1   parallel_4.4.2     scales_1.4.0       yaml_2.3.10
##  [9] fastmap_1.2.0      R6_2.6.1           labeling_0.4.3     generics_0.1.4
## [13] curl_6.4.0         knitr_1.50         pillar_1.11.0      RColorBrewer_1.1-3
## [17] tzdb_0.5.0         rlang_1.1.6        stringi_1.8.7      xfun_0.52
## [21] bit64_4.6.0-1      timechange_0.3.0   cli_3.6.5          withr_3.0.2
## [25] magrittr_2.0.3     digest_0.6.37      grid_4.4.2         vroom_1.6.5
## [29] rstudioapi_0.17.1  hms_1.1.3          lifecycle_1.0.4    vctrs_0.6.5
## [33] evaluate_1.0.4     glue_1.8.0         farver_2.1.2       rmarkdown_2.29
## [37] tools_4.4.2        pkgconfig_2.0.3    htmltools_0.5.8.1
```