Changes on Reading Behaviour for Dutch and Flemish Books during the Lockdown Period

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

This notebook was developed by students of the Online Data Collection Management course integrated into the Master of Marketing Analytics at Tilburg University. Throughout the process we web scraped the "Netherlands & Flemish" group from Goodreads.com from 2019 until October 2021 to study the impact the Covid-19 lockdown had on reading havits for Dutch and Flemish author's. The notebook provides basic statistical inferences regarding our publicly available data set, for a better understanding of it and simultaneously the answer to our research question: - Did the lockdown period had an impact on reading behavior for Dutch and Flemish books? - Did the number of active users increase or decreased? - Did the number of books read by active users increased or decreased? - Did the variety of author's increase or decreased?

We have two possible variables to base teh analysis of our research: 1) data.added: tell us when the user started the book 2) data.read: tell us when the user finished a book

Firstly, we conduct an analysis for data.added and then for data.read with the purpose to identify possible problems regarding the usage of goodreads.com tools (e.g add books who the user actually started reading vs. mark interesting books to read on the future). Another source of possible mistake is the fact users can easily select with one click when they start to read a book, however stating when they actually finished it implies dedicating a certain amount of effort/time to answer goodreads.com questions to enter on the "book.read" user shelf.

Overall to answer our research questions we use 2019 as control year and 2020 as treatment year. The year of 2021 is incomplete since we web scraped from January until the second week of October, where we can see if user behaviour is returning back to normal or not (regarding 2019)

1. Data Preparation and Exploration

```
# Clean environment and make sure all outputs are printed-----------------------
rm(list = ls(all = TRUE))
options(max.print = 999999)

# Load packages-----------------------------------------------------------------
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## 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(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.0.5
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.5
```

```
##
## Attaching package: 'lubridate'
```

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

```
# library(rstudioapi)
# library(haven)
library(summarytools)
```

```
## Warning: package 'summarytools' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```
## For best results, restart R session and update pander using devtools:: or remotes::install_github('ra
```

```
## Load data ---------------------------------------------------------------
setwd("C:/Users/...)
goodreads <- read.csv("data/file1.csv")
```

```
## Request basic statistics------------------------------------------------
summary(goodreads)
```

```
##    reader.id          book.title        author_name          book.url
##  Min.   :    27780   Length:55843       Length:55843       Length:55843
##  1st Qu.: 17004114   Class :character   Class :character   Class :character
##  Median : 44655259   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 51543057
##  3rd Qu.: 76701015
##  Max.   :140899995
##   date.read          date.added        average.rating   user.rating
##  Length:55843       Length:55843       Min.   :0.000   Min.   :0
##  Class :character   Class :character   1st Qu.:3.660   1st Qu.:0
##  Mode  :character   Mode  :character   Median :3.930   Median :0
##                                        Mean   :3.855   Mean   :0
##                                        3rd Qu.:4.140   3rd Qu.:0
##                                        Max.   :5.000   Max.   :0
```

```
initial_row <- nrow(goodreads)
```

Overall the data set has 55.843 observations, eight variables, which six of them are categorical and two are numeric. There ar no duplicates in the data set and the variable user.rating is a straight liner since the standard deviation assumes a value of 0. Such information translates into the fact the users from this group did not provide any rating to the books read. For the research period in cause there are only 685 active members

```
# Request variable information-------------------------------------------------
dim(goodreads)
```

```
## [1] 55843     8
```

```
nrow(goodreads[duplicated(goodreads), ])
```

```
## [1] 0
```

```
dfSummary(goodreads)
```

```
## Data Frame Summary
## goodreads
## Dimensions: 55843 x 8
## Duplicates: 0
##
## ------------------------------------------------------------------------------------------------------------
## No   Variable          Stats / Values                     Freqs (% of Valid)    Graph
## ---- ----------------- ---------------------------------- --------------------- ----------------------
## 1    reader.id         Mean (sd) : 51543057 (39146591)    685 distinct values   . :
##      [integer]         min < med < max:                                         : :   :
##                        27780 < 44655259 < 140899995                             : : . :
##                        IQR (CV) : 59696901 (0.8)                                : : : : :   . . .
##                                                                                 : : : : : : : : : : :
##
## 2    book.title        1. Grand Hotel Europa                 78 ( 0.1%)
##      [character]       2. De meeste mensen deugen:           76 ( 0.1%)
##                        3. De avond is ongemak                60 ( 0.1%)
```

```
##                       4. Animal Farm                      48 ( 0.1%)
##                       5. Het moois dat we delen           48 ( 0.1%)
##                       6. Otmars zonen                     48 ( 0.1%)
##                       7. 1984                             45 ( 0.1%)
##                       8. Ik ben er niet                   44 ( 0.1%)
##                       9. Sapiens: A Brief History         44 ( 0.1%)
##                       10. 't Hooge Nest                   43 ( 0.1%)
##                       [ 34883 others ]                 55309 (99.0%)         IIIIIIIIIIIIIIIIIII
##
## 3    author_name      1. Riley, Lucinda                  405 ( 0.7%)
##      [character]      2. Rowling, J.K.                   350 ( 0.6%)
##                       3. King, Stephen                   330 ( 0.6%)
##                       4. French, Nicci                   211 ( 0.4%)
##                       5. Murakami, Haruki                201 ( 0.4%)
##                       6. Arlidge, M.J.                   172 ( 0.3%)
##                       7. Maas, Sarah J.                  162 ( 0.3%)
##                       8. Atwood, Margaret                154 ( 0.3%)
##                       9. Pfeijffer, Ilja Leonard         151 ( 0.3%)
##                       10. Vandersteen, Willy             145 ( 0.3%)
##                       [ 16329 others ]                 53562 (95.9%)         IIIIIIIIIIIIIIIIII
##
## 4    book.url         1. https://www.goodreads.com        76 ( 0.1%)
##      [character]      2. https://www.goodreads.com        74 ( 0.1%)
##                       3. https://www.goodreads.com        51 ( 0.1%)
##                       4. https://www.goodreads.com        48 ( 0.1%)
##                       5. https://www.goodreads.com        44 ( 0.1%)
##                       6. https://www.goodreads.com        44 ( 0.1%)
##                       7. https://www.goodreads.com        41 ( 0.1%)
##                       8. https://www.goodreads.com        40 ( 0.1%)
##                       9. https://www.goodreads.com        40 ( 0.1%)
##                       10. https://www.goodreads.com       39 ( 0.1%)
##                       [ 37631 others ]                 55346 (99.1%)         IIIIIIIIIIIIIIIIII
##
## 5    date.read        1. 0                             16130 (28.9%)         IIIII
##      [character]      2. Jan-20                          428 ( 0.8%)
##                       3. Jan-19                          277 ( 0.5%)
##                       4. Jan-21                          174 ( 0.3%)
##                       5. Dec 23, 2020                     87 ( 0.2%)
##                       6. Jan-17                           83 ( 0.1%)
##                       7. Jan 31, 2021                     82 ( 0.1%)
##                       8. Apr 05, 2020                     80 ( 0.1%)
##                       9. Apr 04, 2020                     76 ( 0.1%)
##                       10. Feb 07, 2021                    73 ( 0.1%)
##                       [ 1961 others ]                  38353 (68.7%)         IIIIIIIIIIII
##
## 6    date.added       1. Jan 05, 2019                    456 ( 0.8%)
##      [character]      2. Jan 04, 2019                    343 ( 0.6%)
##                       3. May 20, 2020                    343 ( 0.6%)
##                       4. Jun 15, 2019                    312 ( 0.6%)
##                       5. Jan 03, 2019                    297 ( 0.5%)
##                       6. Mar 27, 2020                    278 ( 0.5%)
##                       7. Jan 04, 2020                    257 ( 0.5%)
##                       8. Apr 29, 2020                    256 ( 0.5%)
##                       9. Mar 28, 2020                    251 ( 0.4%)
```

```
##                         10. Aug 26, 2021                    246 ( 0.4%)
##                         [ 1007 others ]                   52804 (94.6%)         IIIIIIIIIIIIIIIII
##
## 7    average.rating     Mean (sd) : 3.9 (0.5)             288 distinct values                 :
##      [numeric]          min < med < max:                                                      : :
##                         0 < 3.9 < 5                                                           : :
##                         IQR (CV) : 0.5 (0.1)                                                . : :
##                                                                                           . : : : .
##
## 8    user.rating        1 distinct value                  0 : 55843 (100.0%)    IIIIIIIIIIIIIIIIII
##      [integer]
## ------------------------------------------------------------------------------------------------------
```

2.Data Cleaning: Unit non-response, Duplicates and Outliers

Firstly, we remove the variable "user.rating" since it is a straight liner Secondly, we transform variables "date.read" and "date.added" into dates in a "yyyy-mm-dd" format since we plan to study changes on readings habits at the weekly level

```
goodreads <- subset(goodreads, select = -user.rating)
goodreads$date_added <- mdy(goodreads$date.added)
goodreads$date_read <- mdy(goodreads$date.read)
```

```
## Warning: 19647 failed to parse.
```

To avoid overlap of variables, the non-formatted date variables are eliminated

```
goodreads <- subset(goodreads, select = -c(date.added, date.read))
```

There are 19.647 values that don't follow the date patterns we applied. The gaps occur on the "user.read" variable, either because the reader did not provide a value (date.read = 0) or because it's missing information regarding the day (date.read = Jan 2021). From the 19.647 books without complete information, it's possible to make the the following distinction: - 16.130 assigned date.read = 0, so probably users that never finished the book or forget to upload such information

-3.517 users only provided the month and the year, without any mention of the exact day when the book was finished

All the 19.647 books were removed. There are no duplicates

```
goodreads <- goodreads %>%
  filter(as.character(date_read) != "NA")


nrow(goodreads[duplicated(goodreads), ])
```

```
## [1] 0
```

Finally we make sure all our variables are in the right format

```
goodreads$date_added <- as.Date(goodreads$date_added, format = "%Y-%m-%d")
goodreads$date_read <- as.Date(goodreads$date_read, format = "%Y-%m-%d")
goodreads$reader.id <- as.factor(goodreads$reader.id)
goodreads$book.url <- as.factor(goodreads$book.url)
goodreads <- data.frame(goodreads)
```

Compare 2019, 2020 and 2021. For that we need to distinguish the different years and weeks. Therefore we assign all users that started to read a book in 2019 with a categorical value of "2019"; for 2020 with a categorical value of "2020" and for 2021 with a value of "2021". We remove week 53 from 2019 and 2020 since it's an incomplete week (28/12-31/12) Because Goodreads.com data was web scraped on a Thursday we remove week 41 from the year 2021 since it is also an incomplete week

```
goodreads <- goodreads %>%
  mutate(year = case_when(
    date_added >= "2019-01-01" & date_added <= "2019-12-31" ~ 2019,
    date_added >= "2020-01-01" & date_added <= "2020-12-31" ~ 2020,
    date_added >= "2021-01-01" ~ 2021
  ))

goodreads$year <- factor(goodreads$year)
goodreads$week_of_year <- week(goodreads$date_added)
goodreads <- goodreads[!(goodreads$week_of_year == 53), ]
goodreads <- goodreads[!(goodreads$week_of_year == 41 & goodreads$year == 2021),]

plot3 <- goodreads %>%
  group_by(year, week_of_year) %>%
  select(year, week_of_year, book.url) %>%
  summarise(total_books = n())
```

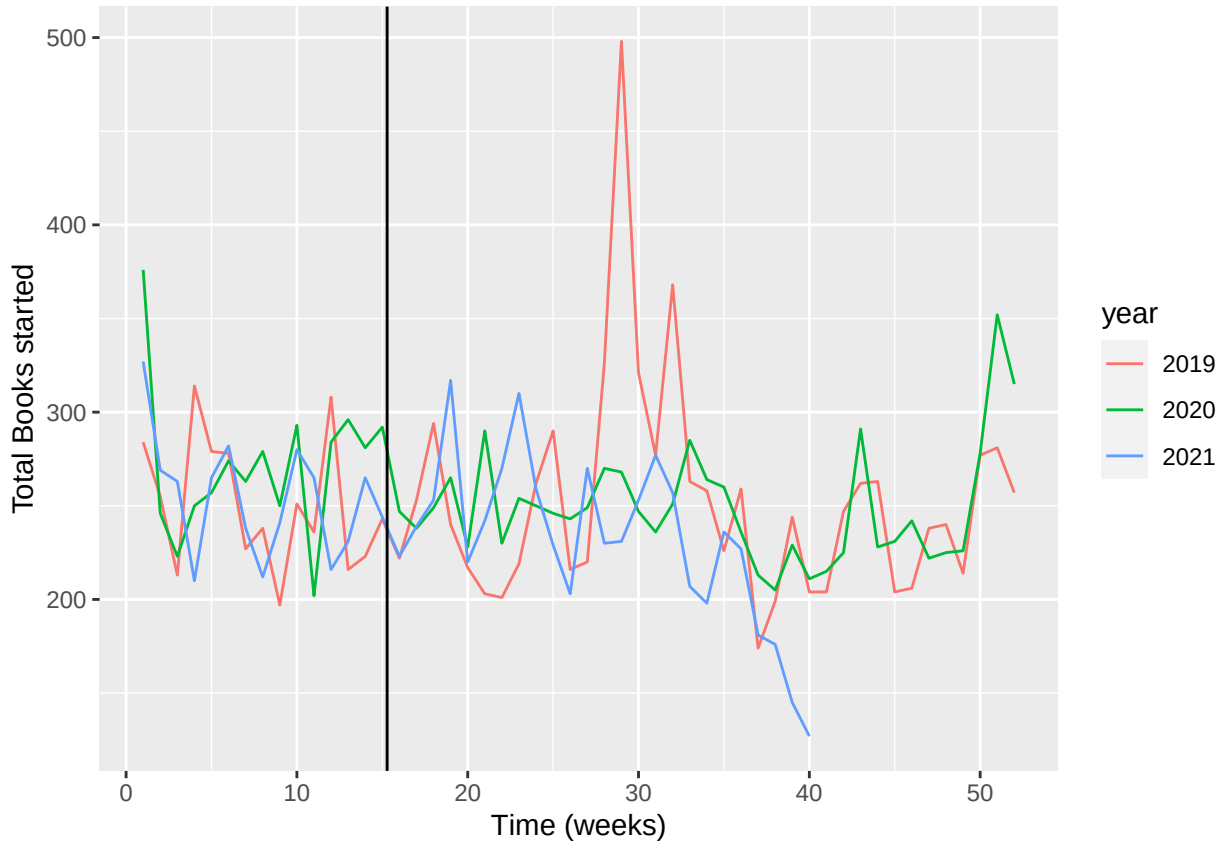
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

The World Health Organization declared a worldwide pandemic state on 11th March 2020. This research uses such date as reference, represented by a black vertical line on the line and bar plots.

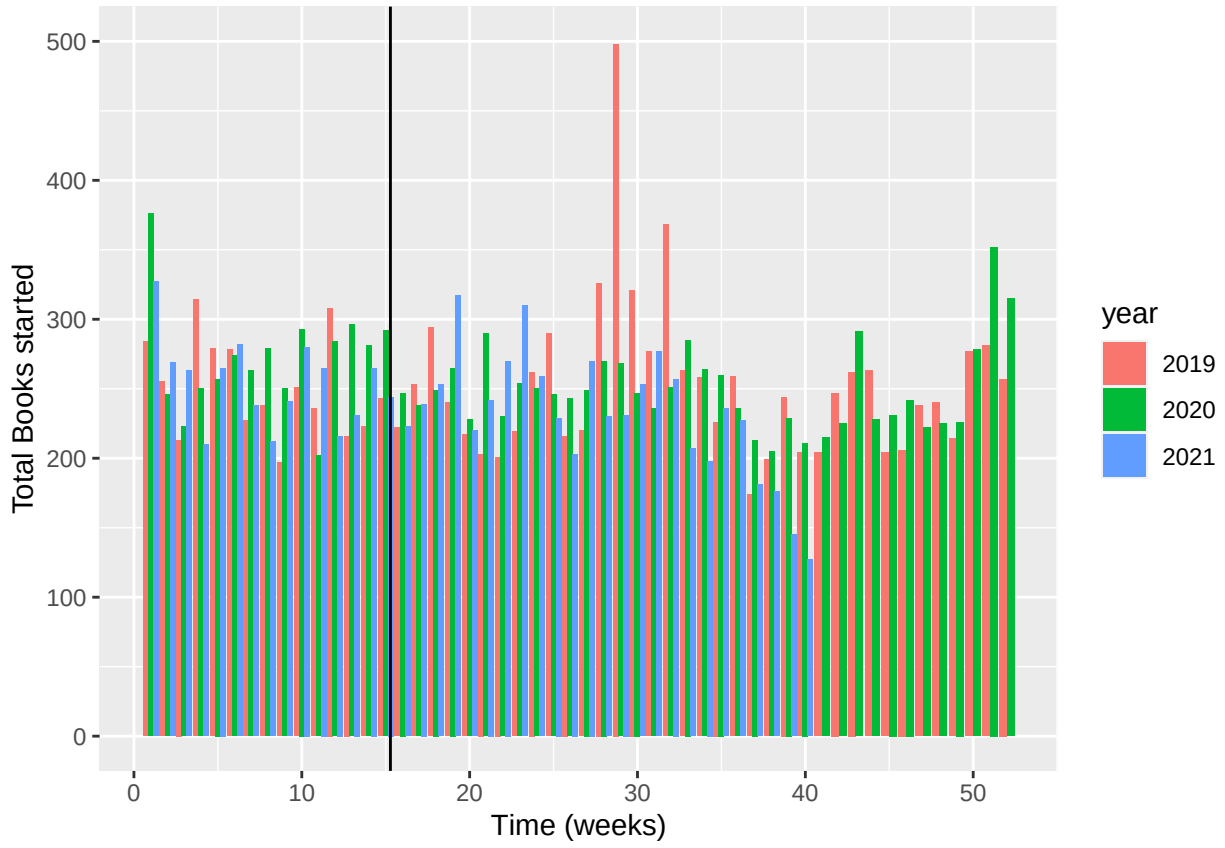
```
who_declaration <- c(seq(as.Date("2020-03-11"), as.Date("2021-10-14"),
  by = "1 day"
))
```

Make plots to visualize the research periods

```
ggplot(data = plot3, aes(x = week_of_year, y = total_books, colour = year)) +
  geom_line() +
  geom_vline(xintercept = 15.282) +
  labs(y = "Total Books started", x = "Time (weeks)")
```




```
ggplot(data = plot3, aes(x = week_of_year, y = total_books, fill = year)) +
  geom_bar(stat = "identity", position = "dodge") +
  ylim(0, 500) +
  geom_vline(xintercept = 15.282) +
  labs(y = "Total Books started", x = "Time (weeks)")
```



The year of 2019 presents 2 abnormal picks. Therefore we explore the presence of possible outliers

```
outliers <- goodreads %>%
  group_by(year, week_of_year, reader.id) %>%
  select(year, week_of_year, reader.id) %>%
  summarise(total_books = n()) %>%
  arrange(desc(total_books))
```

```
## `summarise()` has grouped output by 'year', 'week_of_year'. You can override using the `.groups` argu
```

```
head(outliers)
```

```
## # A tibble: 6 x 4
## # Groups:   year, week_of_year [6]
##   year  week_of_year reader.id total_books
##   <fct>        <dbl> <fct>           <int>
## 1 2019            29 21669112          233
## 2 2019            12 55654304          104
## 3 2019            28 21669112           87
## 4 2019            32 21669112           77
## 5 2020            43 55654304           72
## 6 2019            50 55654304           70
```

```
print(outliers)
```

```
## # A tibble: 17,353 x 4
## # Groups:   year, week_of_year [144]
##    year  week_of_year reader.id total_books
##    <fct>        <dbl> <fct>           <int>
##  1 2019            29 21669112          233
##  2 2019            12 55654304          104
##  3 2019            28 21669112           87
##  4 2019            32 21669112           77
##  5 2020            43 55654304           72
##  6 2019            50 55654304           70
##  7 2020             6 51453456           66
##  8 2019            30 21669112           65
##  9 2019            36 51453456           62
## 10 2020            14 51453456           59
## # ... with 17,343 more rows
```

Conclude that eight users present impossible values like reading from 34 to 233 books in one week. The user have the following reader.id:

1. 21669112
2. 55654304
3. 5145346
4. 108109945
5. 610564
6. 135777991
7. 27721763
8. 28080834

We associate such behavior with the absence of knowledge regarding the Goodreads tool to mark "future books". Such tool allows users to mark books they found interesting and wish to read in the future. The majority of this abnormal values occur during the summer holidays so a season where people have more free time to explore potential books to read and mark them, unfortunately we believe those eight users mark them wrongly. We will eliminate all the 3.611 values relatively to those eight users

```
with_outliers <- nrow(goodreads)
goodreads <- goodreads %>%
  filter(reader.id != 21669112 & reader.id != 55654304 & reader.id != 5145346 &
    reader.id != 108109945 & reader.id != 610564 & reader.id != 135777991 &
    reader.id != 27721763 & reader.id != 28080834)

without_outliers <- nrow(goodreads)
outliers_removed <- with_outliers - without_outliers
print(outliers_removed)
```

```
## [1] 3611
```

We also identify a pick in December 2020, however it's the result of multiple users increasing the individual number of books started and not single users with abnormal values of books read. We explain such pick because it's Christmas and people could not travel due to social-distance measures, using their free time to read. Simultaneously, children books are part of the "Netherlands & Flemish" assortment, which can explain some high values (e.g 16 books read in one week) since they are short books. So no outliers were identified.

```r
outliers_2020 <- goodreads %>%
  filter(year==2020) %>%
  group_by(week_of_year, reader.id) %>%
  select(year, week_of_year, reader.id) %>%
  summarise(total_books = n()) %>%
  arrange(desc(week_of_year))
```

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

```r
print(outliers_2020)
```

```
## # A tibble: 6,290 x 3
## # Groups:   week_of_year [52]
##    week_of_year reader.id total_books
##           <dbl> <fct>           <int>
##  1           52 1063645             4
##  2           52 1212384             1
##  3           52 2394322             1
##  4           52 2489754             7
##  5           52 3276709             2
##  6           52 3624111             3
##  7           52 3713700             2
##  8           52 3866802             1
##  9           52 4276614             2
## 10           52 4745553             1
## # ... with 6,280 more rows
```

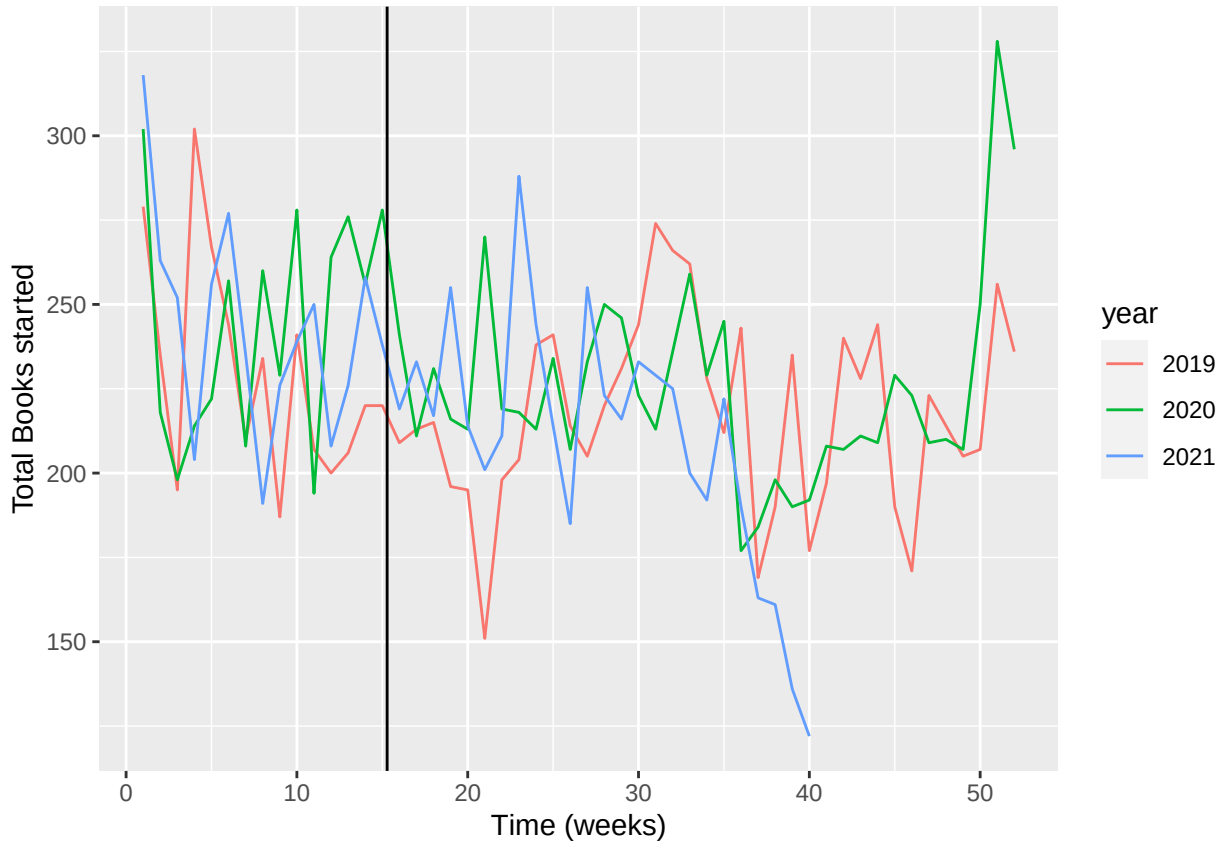
3.Plots: Visualization of the 3 Research Periods, Individually and Together

Make plots to visualize all research periods without the outliers

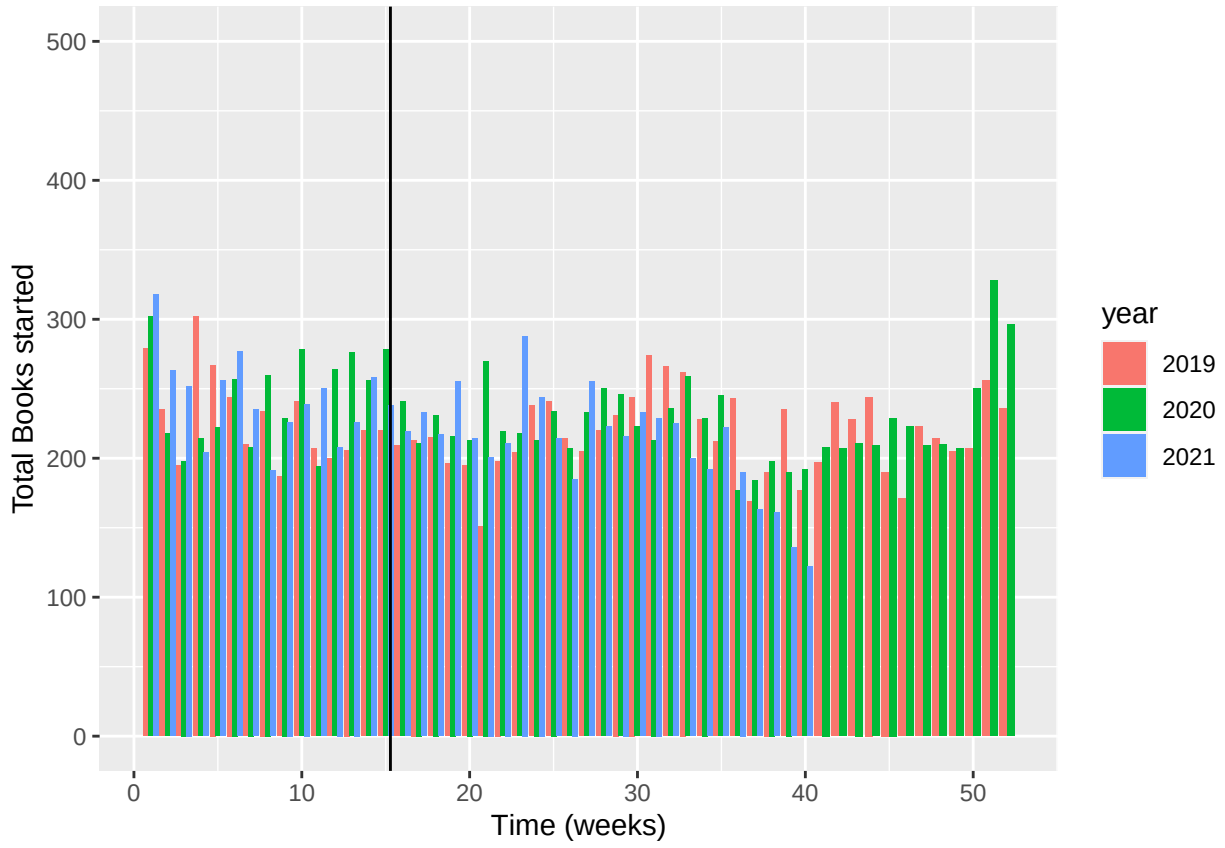
```r
plot3_without_outliers <- goodreads %>%
  filter(week_of_year != 53) %>%
  group_by(year, week_of_year) %>%
  select(year, week_of_year, book.url) %>%
  summarise(total_books = n())
```

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

```r
plot3_without_outliers <- plot3_without_outliers[!(plot3_without_outliers$week_of_year == 41
                                                   & plot3_without_outliers$year == 2021), ]

ggplot(data = plot3_without_outliers, aes(x = week_of_year,
                                          y = total_books, colour = year)) +
  geom_line() +
  geom_vline(xintercept = 15.282) +
  labs(y = "Total Books started", x = "Time (weeks)")
```




```
ggplot(data = plot3_without_outliers, aes(x = week_of_year, y = total_books,
                                          fill = year)) +
  geom_bar(stat = "identity", position = "dodge") +
  ylim(0, 500) +
  geom_vline(xintercept = 15.282) +
  labs(y = "Total Books started", x = "Time (weeks)")
```



Explore the year of 2019, before the outbreak (control year)

```
goodreads_2019 <- goodreads %>%
  filter(grepl("2019", date_added))

total_readers_2019 <- length(unique(goodreads_2019$reader.id))
unique_books2019 <- length(unique(goodreads_2019$book.url))
unique_author2019 <- length(unique(goodreads_2019$author_name))
total_books_2019 <- length(goodreads_2019$year == "2019")

print(total_readers_2019)
```

```
## [1] 449
```

```
print(unique_books2019)
```

```
## [1] 9116
```

```
print(unique_author2019)
```

```
## [1] 4987
```

```r
print(total_books_2019)
```

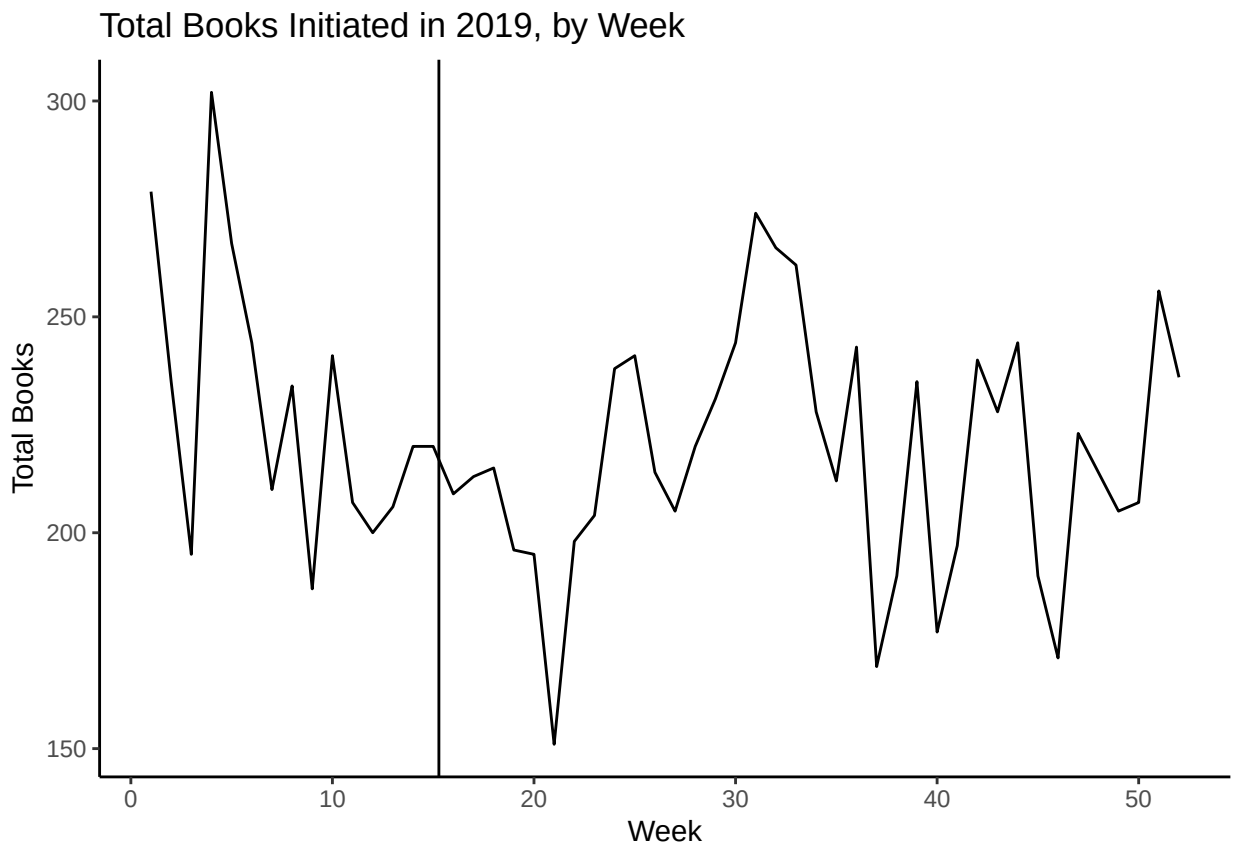
```
## [1] 11488
```

```r
books_per_reader2019 <- unique_books2019 / total_readers_2019
print(books_per_reader2019)
```

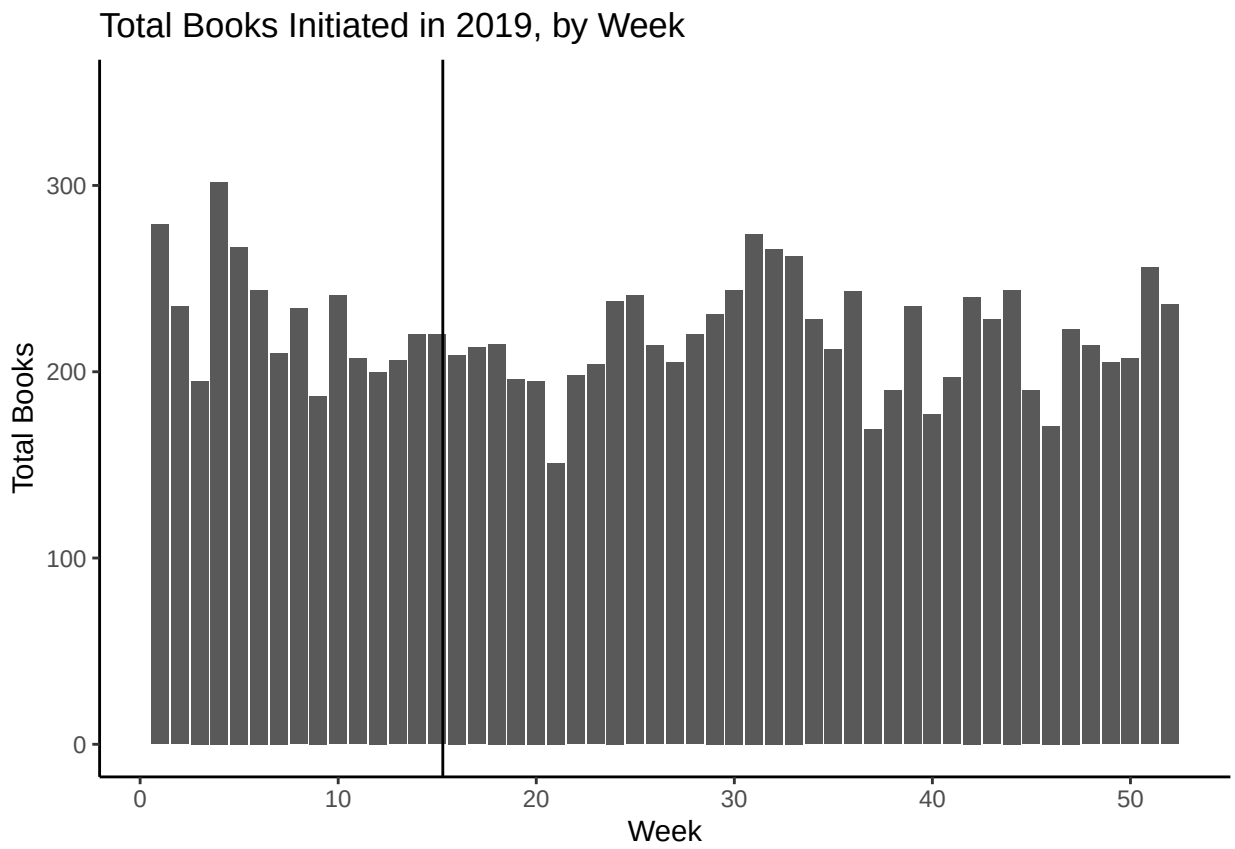
```
## [1] 20.3029
```

Represent the evolution of books started in 2019 in a line and bar plot

```r
goodreads_2019 <- goodreads %>%
  filter(year == 2019)

plot1_2019 <- goodreads_2019 %>%
  group_by(week_of_year) %>%
  select(week_of_year, book.url) %>%
  summarise(total_books = n())


graph_2019_line <- ggplot(plot1_2019, aes(x = week_of_year, y = total_books)) +
  geom_line() +
  xlim(1, 52) +
  theme_bw() +
  geom_vline(xintercept = 15.282) +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Initiated in 2019, by Week")

graph_2019_line
```




```
graph_2019_bar <- ggplot(plot1_2019, aes(x = week_of_year, y = total_books)) +
  geom_bar(stat = "identity") +
  ylim(0, 350) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Initiated in 2019, by Week")

graph_2019_bar
```



Explore the year of 2020 (treatment year)

```
goodreads_2020 <- goodreads %>%
  filter(year == 2020)

total_readers_2020 <- length(unique(goodreads_2020$reader.id))
unique_books2020 <- length(unique(goodreads_2020$book.url))
unique_author2020 <- length(unique(goodreads_2020$author_name))
total_books_2020 <- length(goodreads_2020$year == "2020")

print(total_books_2020)
```

```
## [1] 11989
```

```
print(total_readers_2020)
```

```
## [1] 478
```

```
print(unique_books2020)
```

```
## [1] 9245
```

```r
print(unique_author2020)
```

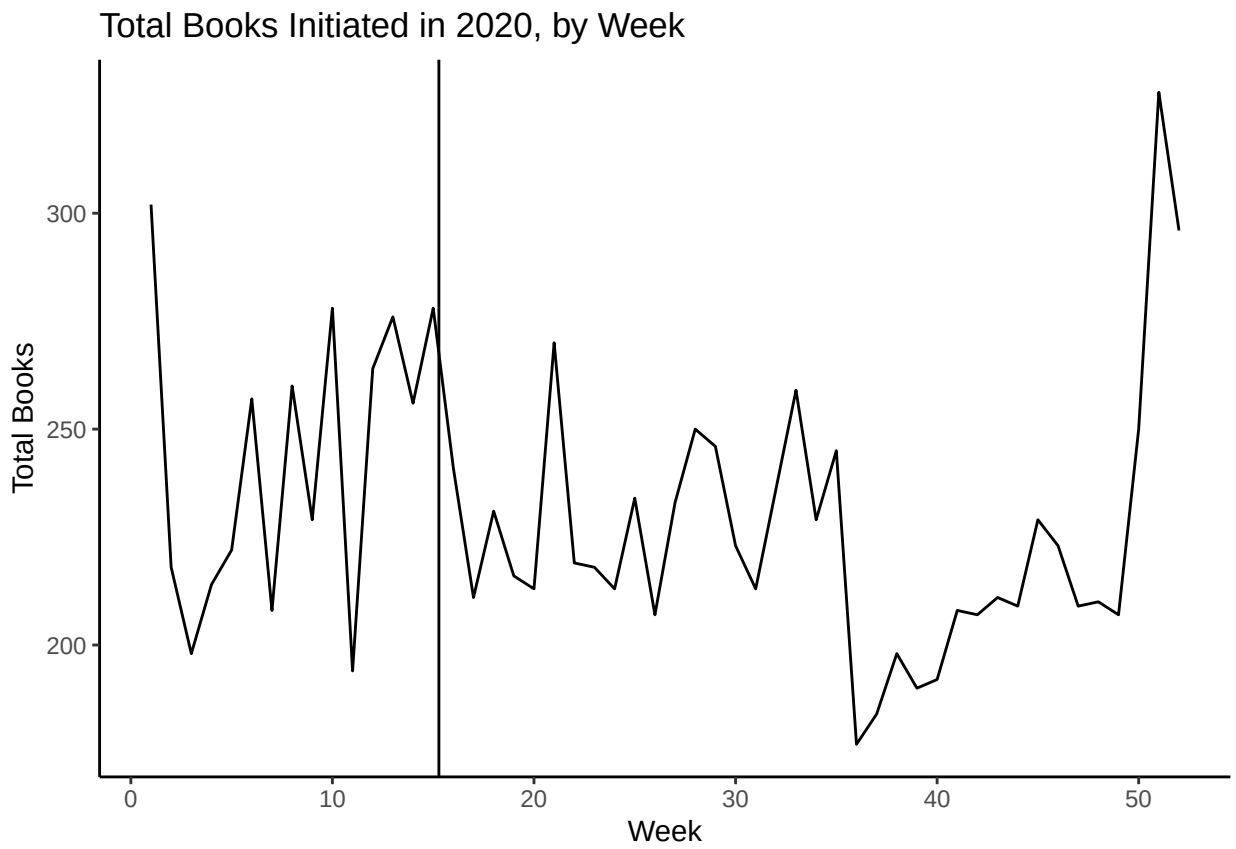
```
## [1] 5070
```

```r
books_per_reader2020 <- unique_books2020 / total_readers_2020
print(books_per_reader2020)
```

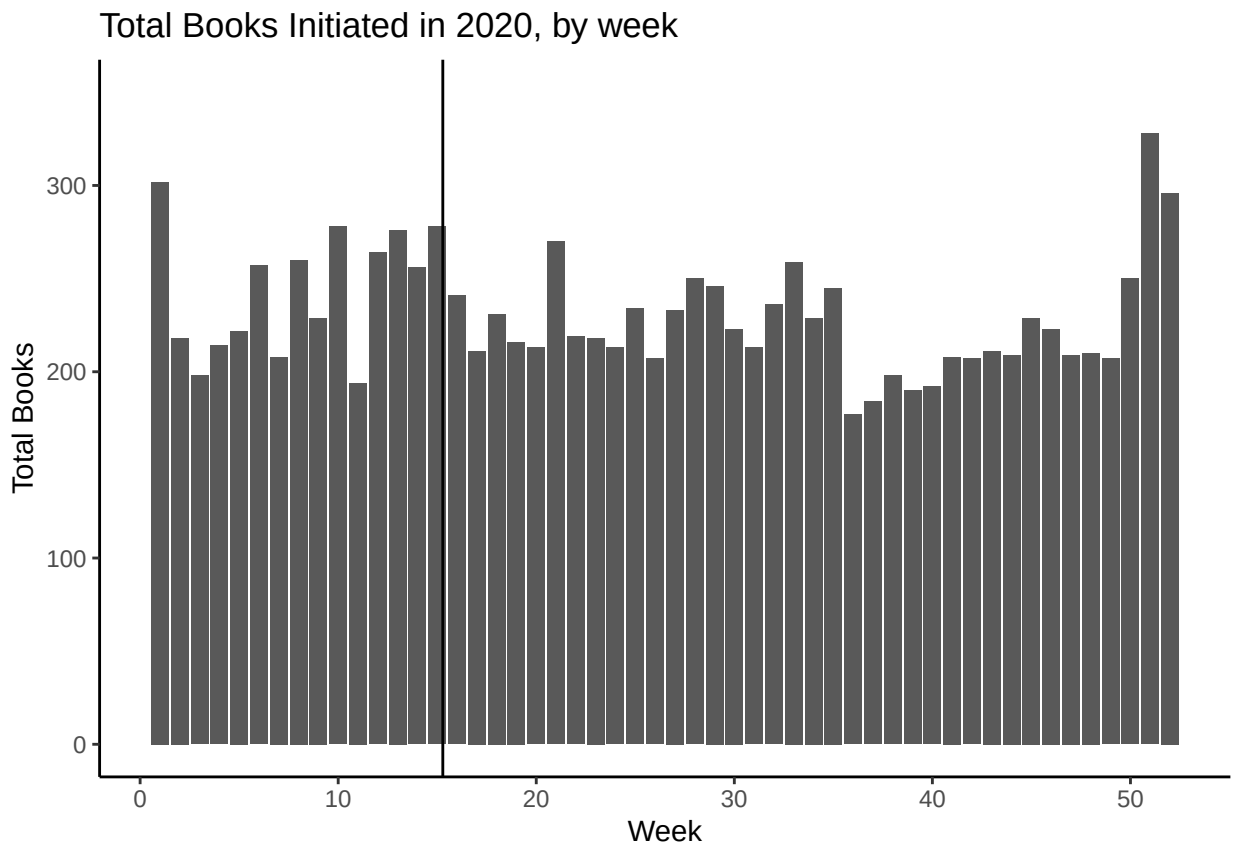
```
## [1] 19.341
```

Represent the evolution of books started in 2020 in a line and bar plot

```r
plot1_2020 <- goodreads_2020 %>%
  group_by(week_of_year) %>%
  select(week_of_year, book.url) %>%
  summarise(total_books = n())



graph_2020_line <- ggplot(plot1_2020, aes(x = week_of_year, y = total_books)) +
  geom_line() +
  xlim(1, 52) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Initiated in 2020, by Week")

graph_2020_line
```




```
ggplot(plot1_2020, aes(x = week_of_year, y = total_books)) +
  geom_bar(stat = "identity", position = position_dodge(width = 1)) +
  ylim(0, 350) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Initiated in 2020, by week")
```



Explore the year of 2021 (incomplete year, only goes until week 40)

```
goodreads_2021 <- goodreads %>%
  filter(year == 2021)

total_readers_2021 <- length(unique(goodreads_2021$reader.id))
unique_books2021 <- length(unique(goodreads_2021$book.url))
unique_author2021 <- length(unique(goodreads_2021$author_name))
total_books_2021 <- length(goodreads_2021$year == "2021")

print(total_books_2021)
```

```
## [1] 8889
```

```
print(total_readers_2021)
```

```
## [1] 450
```

```
print(unique_books2021)
```

```
## [1] 7165
```

```r
print(unique_author2021)
```

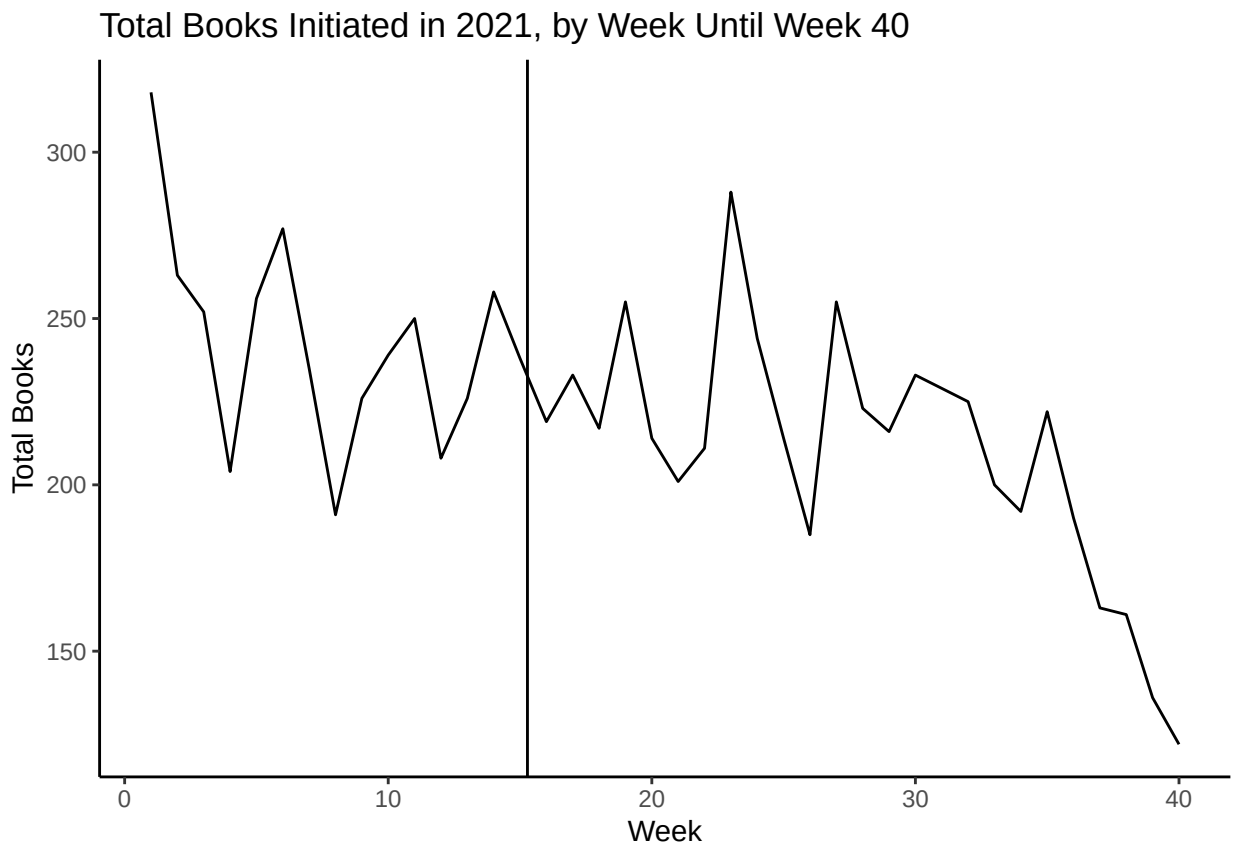
```
## [1] 4287
```

```r
books_per_reader2021 <- unique_books2021 / total_readers_2021
print(books_per_reader2021)
```

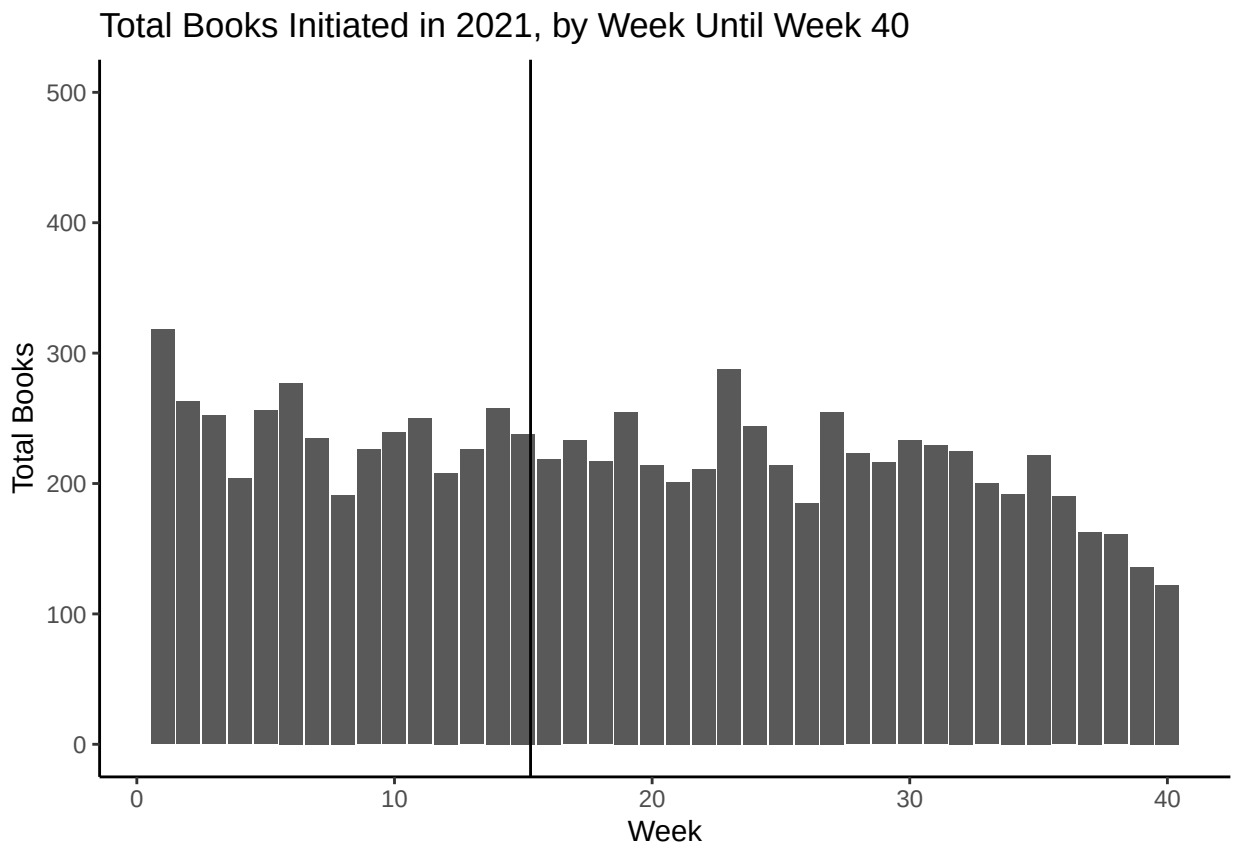
```
## [1] 15.92222
```

Represent the evolution of books started in 2021 in a line and bar plot

```r
plot1_2021 <- goodreads_2021 %>%
  group_by(week_of_year) %>%
  select(week_of_year, book.url) %>%
  summarise(total_books = n())



graph_2021_line <- ggplot(plot1_2021, aes(x = week_of_year, y = total_books)) +
  geom_line() +
  xlim(1, 40) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Initiated in 2021, by Week Until Week 40")

graph_2021_line
```




```
ggplot(plot1_2021, aes(x = week_of_year, y = total_books)) +
  geom_bar(stat = "identity", position = position_dodge(width = 1)) +
  ylim(0, 500) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Initiated in 2021, by Week Until Week 40")
```




```
print(total_readers_2019)
```

```
## [1] 449
```

```
print(unique_books2019)
```

```
## [1] 9116
```

```
print(unique_author2019)
```

```
## [1] 4987
```

```
print(total_books_2019)
```

```
## [1] 11488
```

```
print("----------------")
```

```
## [1] "----------------"
```

```r
print(total_readers_2020)
```

```
## [1] 478
```

```r
print(unique_books2020)
```

```
## [1] 9245
```

```r
print(unique_author2020)
```

```
## [1] 5070
```

```r
print(total_books_2020)
```

```
## [1] 11989
```

```r
print("---------------")
```

```
## [1] "---------------"
```

```r
print(total_readers_2021)
```

```
## [1] 450
```

```r
print(unique_books2021)
```

```
## [1] 7165
```

```r
print(unique_author2021)
```

```
## [1] 4287
```

```r
print(total_books_2021)
```

```
## [1] 8889
```

3.Conclusions

Having in mind our research questions - Did the lockdown period had an impact on reading behavior for Dutch and Flemish books? 1) Did the number of active users increase or decreased? The number of active users increased from 449 to 478 from 2019 to 2020. As social distance measures become less restrictive there is reduction of active users from 2020 to 2021, from 478 to 450. In fact if we compare 2019 with 2021, there is an increase of only one active user. We may infer that lockdown lead to an increase of active users but as life goes back to normal there is an decrease of active users.

2) Did the number of books read by active users increased or decreased? Because the number of active users increased from 2019 to 2020 we expect to see a similar effect on books read. There is an increase from 11.488 books to 11.989 books read from 2019 to 2020. We can not infer anything regarding 2021 since we only web scraped until the second week of October.

3)Did the variety of author's increase or decreased? There is an increase in author's variety from 2019 to 2020. This may suggest that people used there free time not only to read more but to explore new author's. Because 2021 is an incomplete year we can not infer anything.

We think that there could be a problem with the usage of date_added variable. Therefore we conduct a second analysis using date_read, where the focus is when the consumer finished the book, since people will only add books they actually finished, avoid possible mistakes between started and finished books vs. books consumer want to read.

B.SECOND ANALYSIS

B.1. Data Preparation

```
setwd("C:/Users/...")
goodreads2 <- read.csv("data/file1.csv")
```

B.2.Data Cleaning: Unit non-response, Duplicates and Outliers

Firstly, we remove the variable "user.rating" since it is a straight liner Secondly, we transform variables "date.read" and "date.added" into dates in a "yyyy-mm-dd" format since we plan to study changes on readings habits at the weekly level

```
goodreads2 <- subset(goodreads2, select = -user.rating)
goodreads2$date_read <- mdy(goodreads2$date.read)
```

```
## Warning: 19647 failed to parse.
```

To avoid overlap of variables, the non-formatted date variables are eliminated

```
goodreads2 <- subset(goodreads2, select = -c(date.added, date.read))
```

There are 19.647 values that don't follow the date patterns we applied. The gaps occur on the "user.read" variable, either because the reader did not provide a value (date.read = 0) or because it's missing information regarding the day (date.read = Jan 2021). From the 19.647 books without complete information, it's possible to make the the following distinction: - 16.130 assigned date.read = 0, so probably users that never finished the book or forget to upload such information

-3.517 users only provided the month and the year, without any mention of the exact day when the book was finished

All the 19.647 books were removed. There are no duplicates

```
goodreads2 <- goodreads2 %>%
  filter(as.character(date_read) != "NA")
```

```
nrow(goodreads2[duplicated(goodreads2), ])
```

```
## [1] 0
```

Finally we make sure all our variables are in the right format

```
goodreads2$date_read <- as.Date(goodreads2$date_read, format = "%Y-%m-%d")
goodreads2$reader.id <- as.factor(goodreads2$reader.id)
goodreads2$book.url <- as.factor(goodreads2$book.url)
goodreads2 <- data.frame(goodreads2)
```

Compare 2019, 2020 and 2021. For that we need to distinguish the different years and weeks. Therefore we assign all users that started to read a book in 2019 with a categorical value of "2019"; for 2020 with a categorical value of "2020" and for 2021 with a value of "2021" (making sure that each year is perceived as a level, so we factorize the variable). We remove week 53 from 2019 and 2020 since it's an incomplete week (28/12-31/12) Because Goodreads.com data was web-scrapped on a Thursday we also remove week 41 from the year 2021 since it is also an incomplete week. Any possible duplicates and NA's are eliminated.

```
goodreads2 <- goodreads2 %>%
  mutate(year = case_when(
    date_read >= "2019-01-01" & date_read <= "2019-12-31" ~ 2019,
    date_read >= "2020-01-01" & date_read <= "2020-12-31" ~ 2020,
    date_read >= "2021-01-01" ~ 2021
  ))

goodreads2$year <- factor(goodreads2$year)
goodreads2$week_of_year <- week(goodreads2$date_read)
goodreads2 <- goodreads2[!(goodreads2$week_of_year == 53), ]
goodreads2 <- goodreads2[!(goodreads2$week_of_year == 41 &
                             goodreads2$year == 2021), ]

goodreads2 <- goodreads2[!duplicated(goodreads2), ]
goodreads2 <- na.omit(goodreads2)
nrow(goodreads2)
```

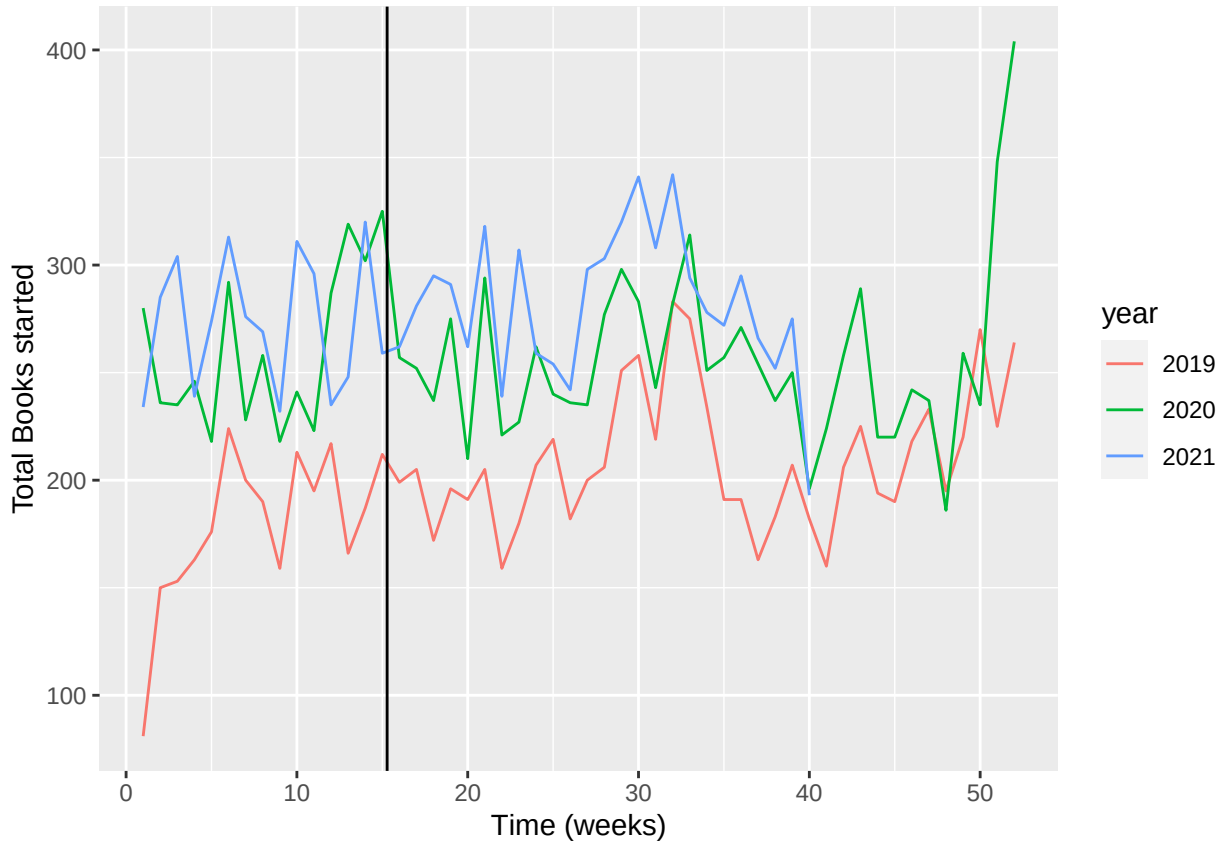
```
## [1] 34975
```

```
plot1_second <- goodreads2 %>%
  group_by(year, week_of_year) %>%
  select(year, week_of_year, book.url) %>%
  summarise(total_books = n())
```

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

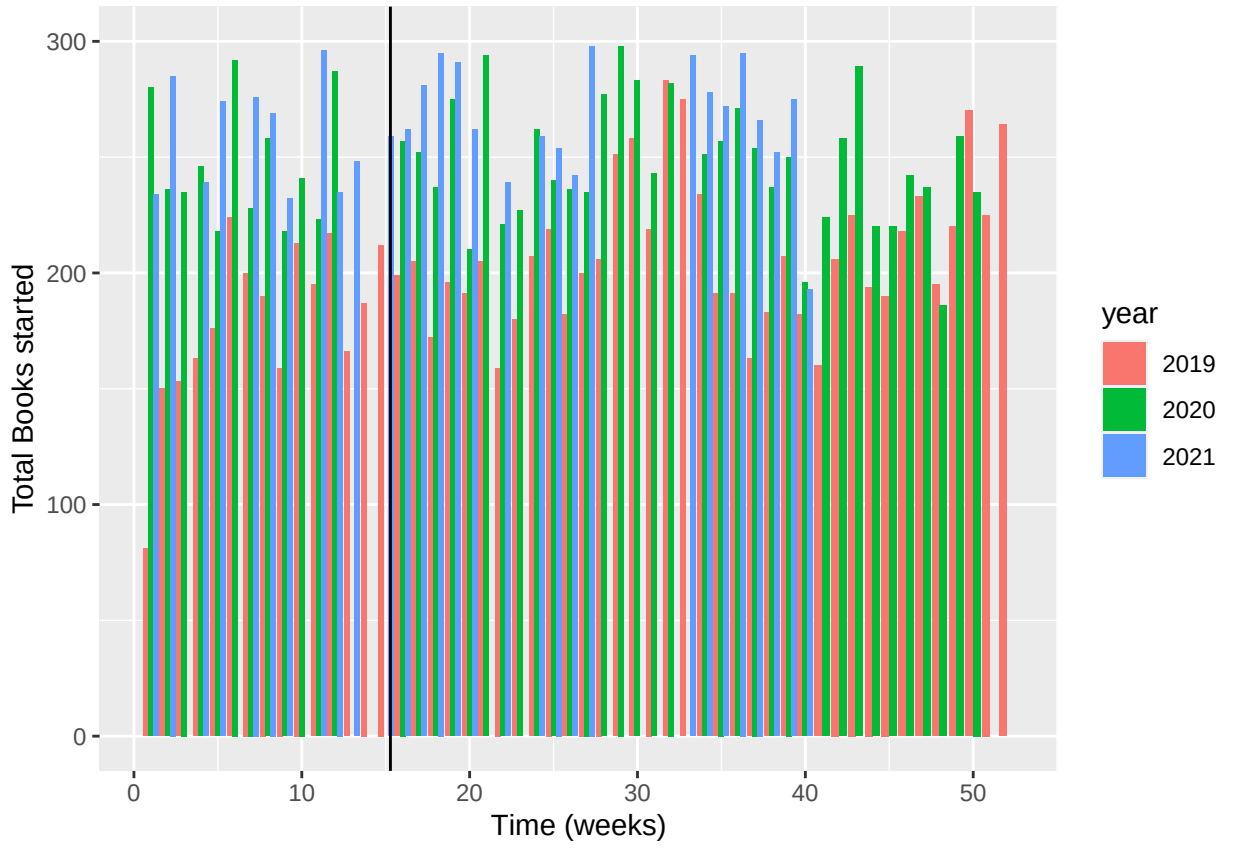
Make plots to visualize the research period

```
ggplot(data = plot1_second, aes(x = week_of_year, y = total_books, colour = year)) +
  geom_line() +
  geom_vline(xintercept = 15.282) +
  labs(y = "Total Books started", x = "Time (weeks)")
```




```
ggplot(data = plot1_second, aes(x = week_of_year, y = total_books, fill = year)) +
  geom_bar(stat = "identity", position = "dodge") +
  ylim(0, 300) +
  geom_vline(xintercept = 15.282) +
  labs(y = "Total Books started", x = "Time (weeks)")
```

```
## Warning: Removed 17 rows containing missing values (geom_bar).
```



The year of 2019 presents 2 abnormal picks. Therefore we explore the presence of possible outliers

```
outliers <- goodreads %>%
  group_by(year, week_of_year, reader.id) %>%
  select(year, week_of_year, reader.id) %>%
  summarise(total_books = n()) %>%
  arrange(desc(total_books))
```

```
## `summarise()` has grouped output by 'year', 'week_of_year'. You can override using the `.groups` argu
```

```
head(outliers)
```

```
## # A tibble: 6 x 4
## # Groups:   year, week_of_year [6]
##   year  week_of_year reader.id total_books
##   <fct>        <dbl> <fct>           <int>
## 1 2020             6 51453456           66
## 2 2019            36 51453456           62
## 3 2020            14 51453456           59
## 4 2020            13 51453456           53
## 5 2019            25 51453456           45
## 6 2020            15 51453456           45
```

```
print(outliers)
```

```
## # A tibble: 16,899 x 4
## # Groups:   year, week_of_year [144]
##    year  week_of_year reader.id total_books
##    <fct>        <dbl> <fct>           <int>
##  1 2020             6 51453456           66
##  2 2019            36 51453456           62
##  3 2020            14 51453456           59
##  4 2020            13 51453456           53
##  5 2019            25 51453456           45
##  6 2020            15 51453456           45
##  7 2019            33 51453456           34
##  8 2020            24 51453456           33
##  9 2019            31 51453456           32
## 10 2019            23 51453456           30
## # ... with 16,889 more rows
```

Conclude that five users present impossible values like reading from 24 to 66 books in one week. The user have the following reader.id:

1. 51453456
2. 54761655
3. 30700335 4.124091570 5.104199220

By looking at the assortment of books we see some are children books. So we may be looking at readers that read between 1/2 books per day during their vacations plus children books to their son's and daughter. We decided to establish a selling of 21 books per week (3 book maximum per day)

We will eliminate all the values relatively to those five users

```
with_outliers_second <- nrow(goodreads2)

goodreads2 <- goodreads2 %>%
  filter(reader.id != 51453456 & reader.id != 54761655 & reader.id != 30700335 &
    reader.id != 124091570 & reader.id != 104199220)

without_outliers_second <- nrow(goodreads2)
outliers_removed_second <- with_outliers_second - without_outliers_second
print(outliers_removed_second)
```

```
## [1] 2600
```

Remove NA's

```
goodreads2 <- na.omit(goodreads2)
nrow(goodreads2)
```

```
## [1] 32375
```

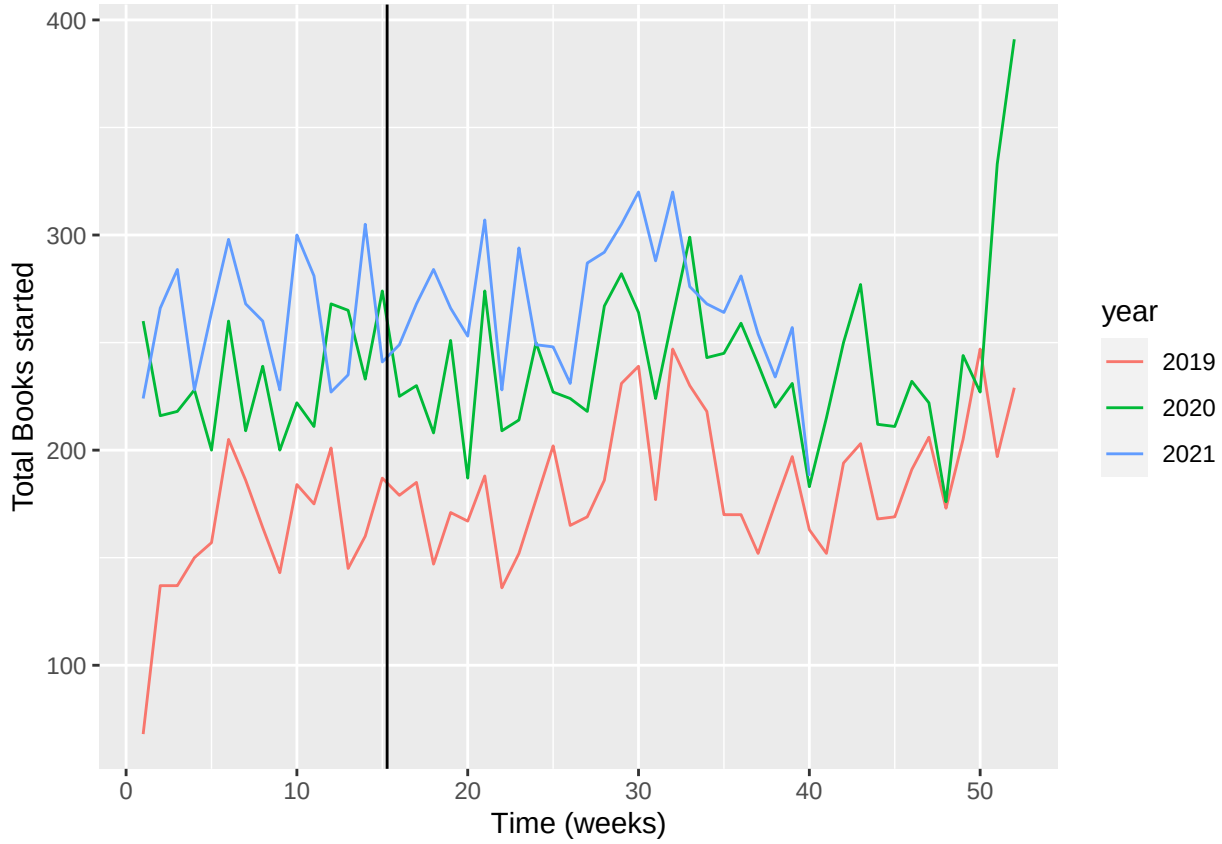
B3.PLOTS: Visualization of the 3 Research Periods, Individually and Together

Make plots to visualize all research periods without the outliers

```
plot3_without_outliers_second <- goodreads2 %>%
  filter(week_of_year != 53) %>%
  group_by(year, week_of_year) %>%
  select(year, week_of_year, book.url) %>%
  summarise(total_books = n())
```

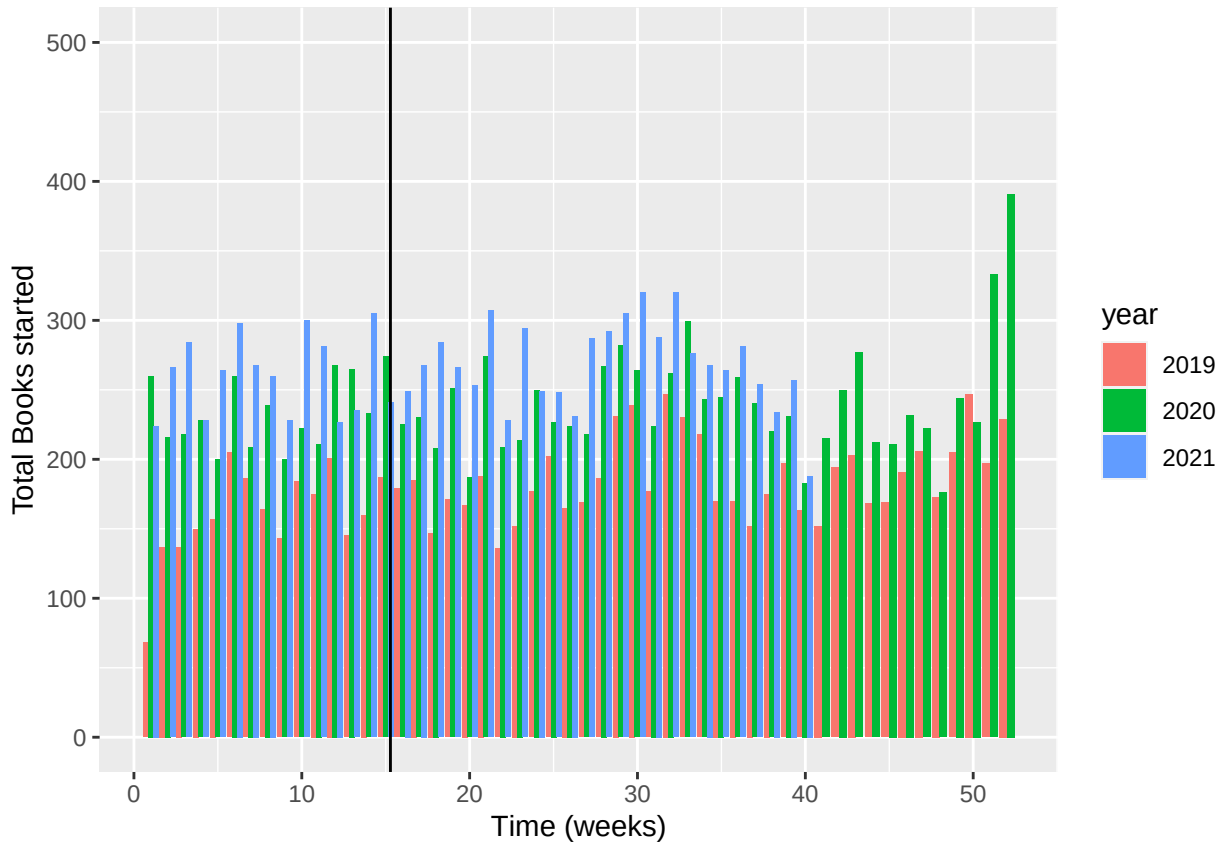
```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
plot3_without_outliers_second <- plot3_without_outliers_second[!
                                (plot3_without_outliers_second$week_of_year == 41
                                 & plot3_without_outliers_second$year == 2021), ]

ggplot(data = plot3_without_outliers_second, aes(x = week_of_year, y = total_books,
                                                 colour = year)) +
  geom_line() +
  geom_vline(xintercept = 15.282) +
  labs(y = "Total Books started", x = "Time (weeks)")
```




```
ggplot(data = plot3_without_outliers_second, aes(x = week_of_year, y = total_books,
                                                 fill = year)) +
  geom_bar(stat = "identity", position = "dodge") +
  ylim(0, 500) +
```

```
  geom_vline(xintercept = 15.282) +
  labs(y = "Total Books started", x = "Time (weeks)")
```



Explore the year of 2019, before the outbreak (control year)

```
goodreads_2019_second <- goodreads2 %>%
  filter(grepl("2019", date_read))

total_readers_2019_second <- length(unique(goodreads_2019_second$reader.id))
unique_books2019_second <- length(unique(goodreads_2019_second$book.url))
unique_author2019_second <- length(unique(goodreads_2019_second$author_name))
total_books2019_second <- length(goodreads_2019_second$year == "2019")

print(total_books2019_second)
```

```
## [1] 9326
```

```
print(total_readers_2019_second)
```

```
## [1] 437
```

```
print(unique_books2019_second)
```

```
## [1] 7863
```

```r
print(unique_author2019_second)
```

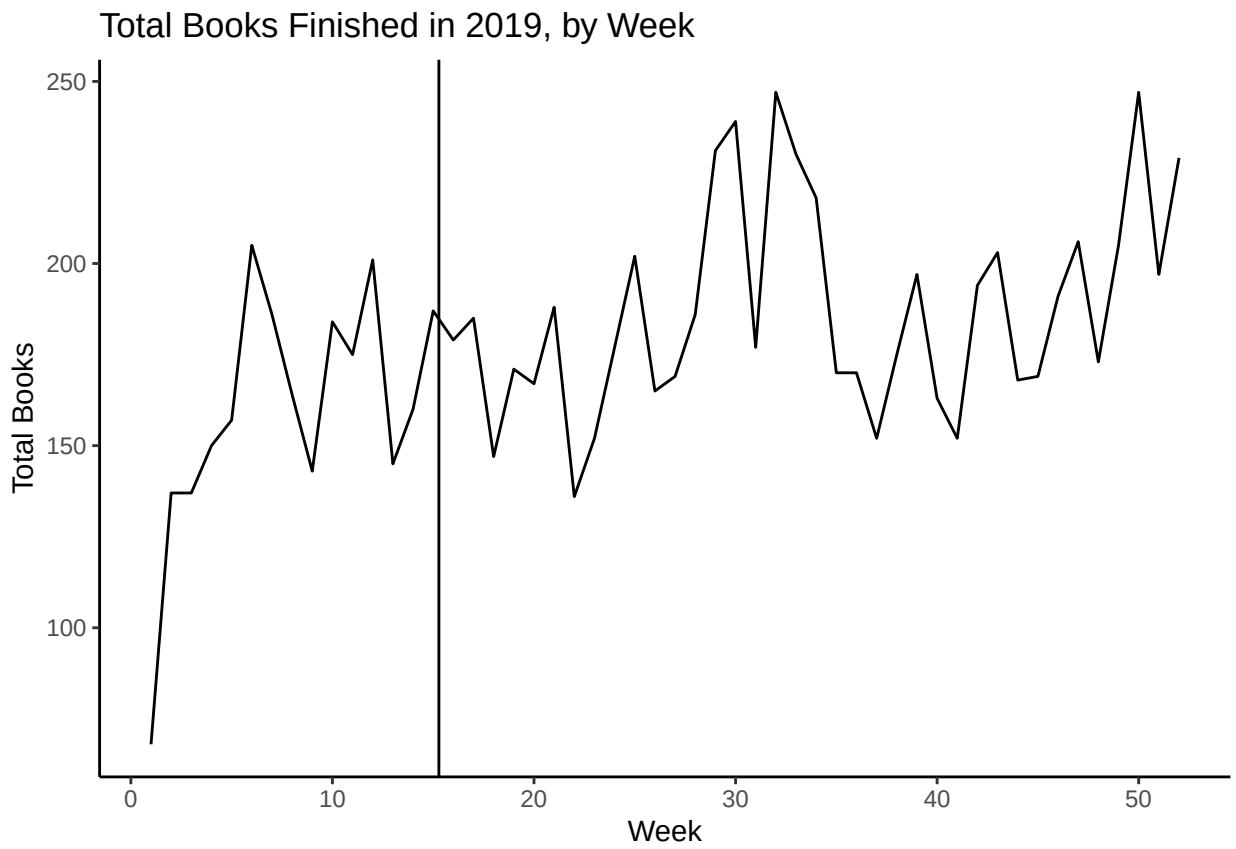
```
## [1] 4346
```

```r
books_per_reader2019_second <- unique_books2019_second / total_readers_2019_second
print(books_per_reader2019_second)
```

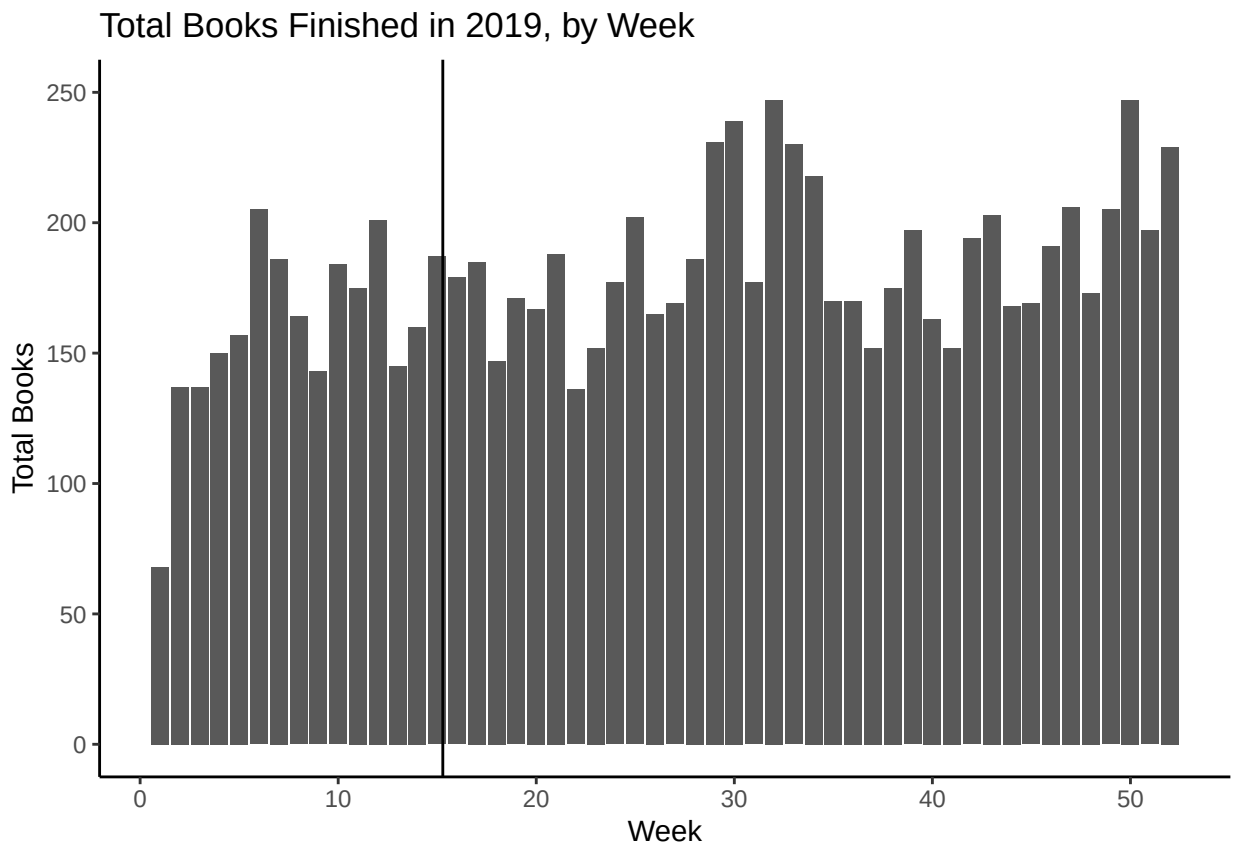
```
## [1] 17.99314
```

Represent the evolution of books started in 2019 in a line and bar plot

```r
goodreads_2019_second <- goodreads2 %>%
  filter(year == 2019)

plot1_2019_second <- goodreads_2019_second %>%
  group_by(week_of_year) %>%
  select(week_of_year, book.url) %>%
  summarise(total_books = n())


graph_2019_line_second <- ggplot(plot1_2019_second, aes(x = week_of_year,
                                                        y = total_books)) +
  geom_line() +
  xlim(1, 52) +
  theme_bw() +
  geom_vline(xintercept = 15.282) +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Finished in 2019, by Week")

graph_2019_line_second
```




```
graph_2019_bar_second <- ggplot(plot1_2019_second, aes(x = week_of_year,
                                                       y = total_books)) +
  geom_bar(stat = "identity") +
  ylim(0, 250) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Finished in 2019, by Week")

graph_2019_bar_second
```



Explore the year of 2020 (treatment year)

```
goodreads_2020_second <- goodreads2 %>%
  filter(year == 2020)

total_readers_2020_second <- length(unique(goodreads_2020_second$reader.id))
unique_books2020_second <- length(unique(goodreads_2020_second$book.url))
unique_author2020_second <- length(unique(goodreads_2020_second$author_name))
total_books2020_second <- length(goodreads_2020_second$year == "2020")

print(total_books2020_second)
```

```
## [1] 12429
```

```
print(total_readers_2020_second)
```

```
## [1] 496
```

```
print(unique_books2020_second)
```

```
## [1] 9871
```

```r
print(unique_author2020_second)
```

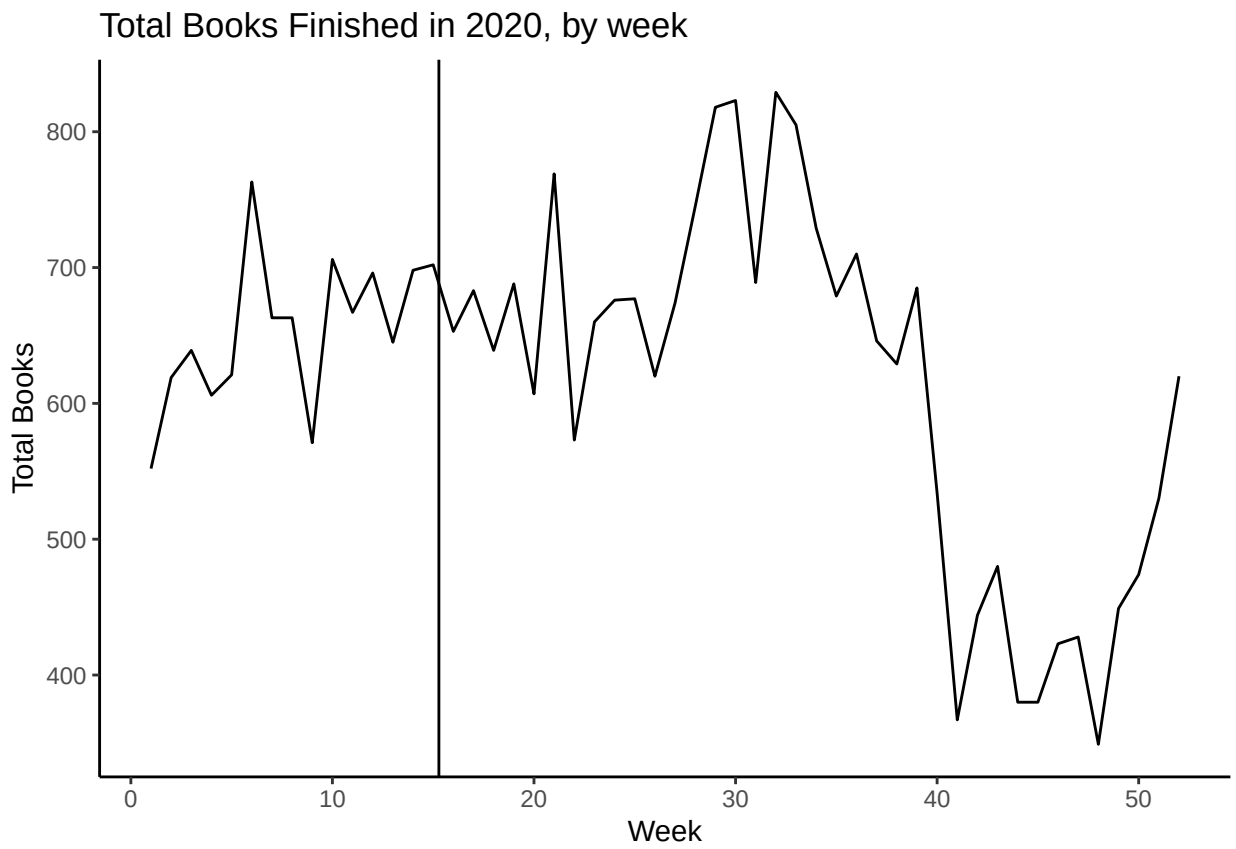
```
## [1] 5416
```

```r
books_per_reader2020_second <- unique_books2020_second / total_readers_2020_second
print(books_per_reader2020_second)
```

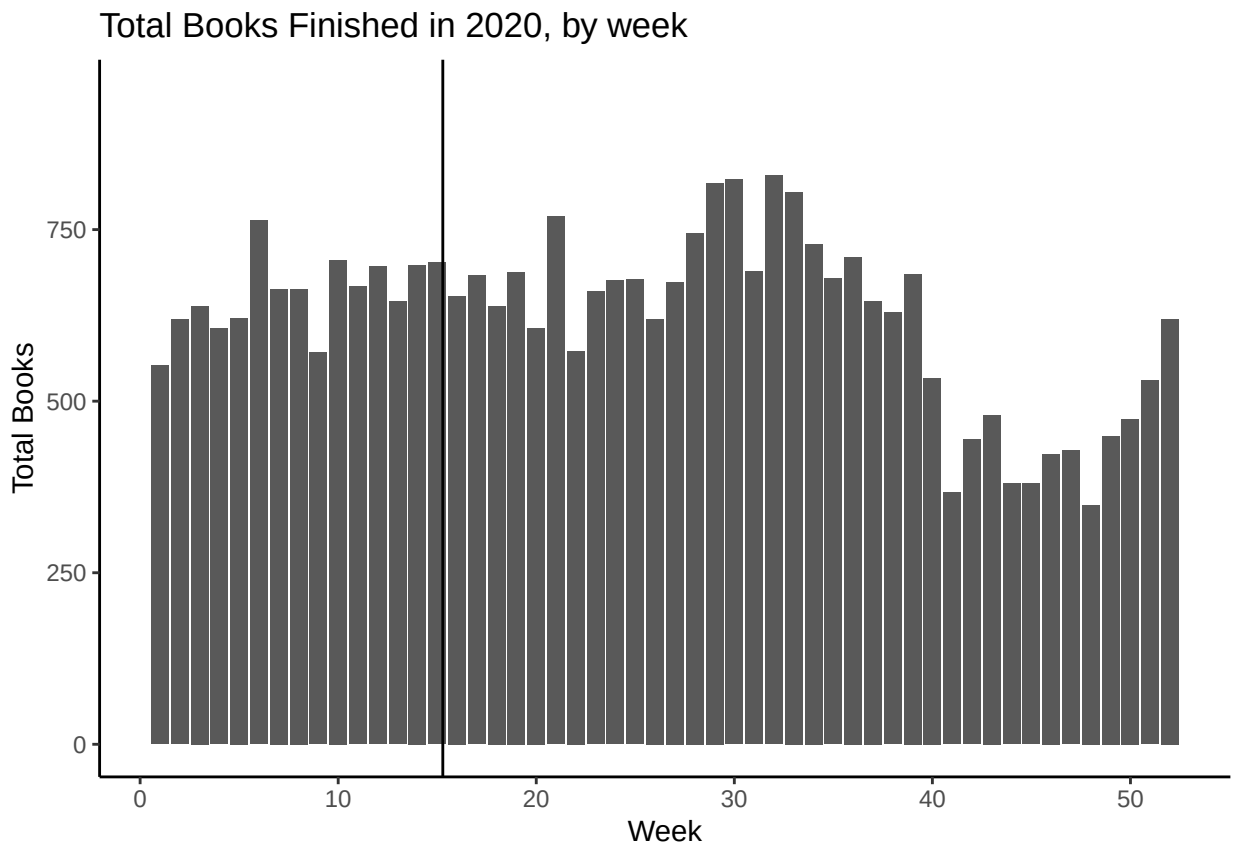
```
## [1] 19.90121
```

Represent the evolution of books started in 2020 in a line and bar plot

```r
plot1_2020_second <- goodreads2 %>%
  group_by(week_of_year) %>%
  select(week_of_year, book.url) %>%
  summarise(total_books = n())



graph_2020_line_second <- ggplot(plot1_2020_second, aes(x = week_of_year,
                                                        y = total_books)) +
  geom_line() +
  xlim(1, 52) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Finished in 2020, by week")

graph_2020_line_second
```




```
graph_2020_bar_second <- ggplot(plot1_2020_second, aes(x = week_of_year,
                                                       y = total_books)) +
  geom_bar(stat = "identity", position = position_dodge(width = 1)) +
  ylim(0, 950) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Finished in 2020, by week")


graph_2020_bar_second
```



Explore the year of 20201 (incomplete year, only goes until week 40)

```
goodreads_2021_second <- goodreads2 %>%
  filter(year == 2021)

total_readers_2021_second <- length(unique(goodreads_2021_second$reader.id))
unique_books2021_second <- length(unique(goodreads_2021_second$book.url))
unique_author2021_second <- length(unique(goodreads_2021_second$author_name))
total_books2021_second <- length(goodreads_2021_second$year == "2021")

print(total_books2021_second)
```

```
## [1] 10620
```

```
print(total_readers_2021_second)
```

```
## [1] 473
```

```
print(unique_books2021_second)
```

```
## [1] 8428
```

```r
print(unique_author2021_second)
```

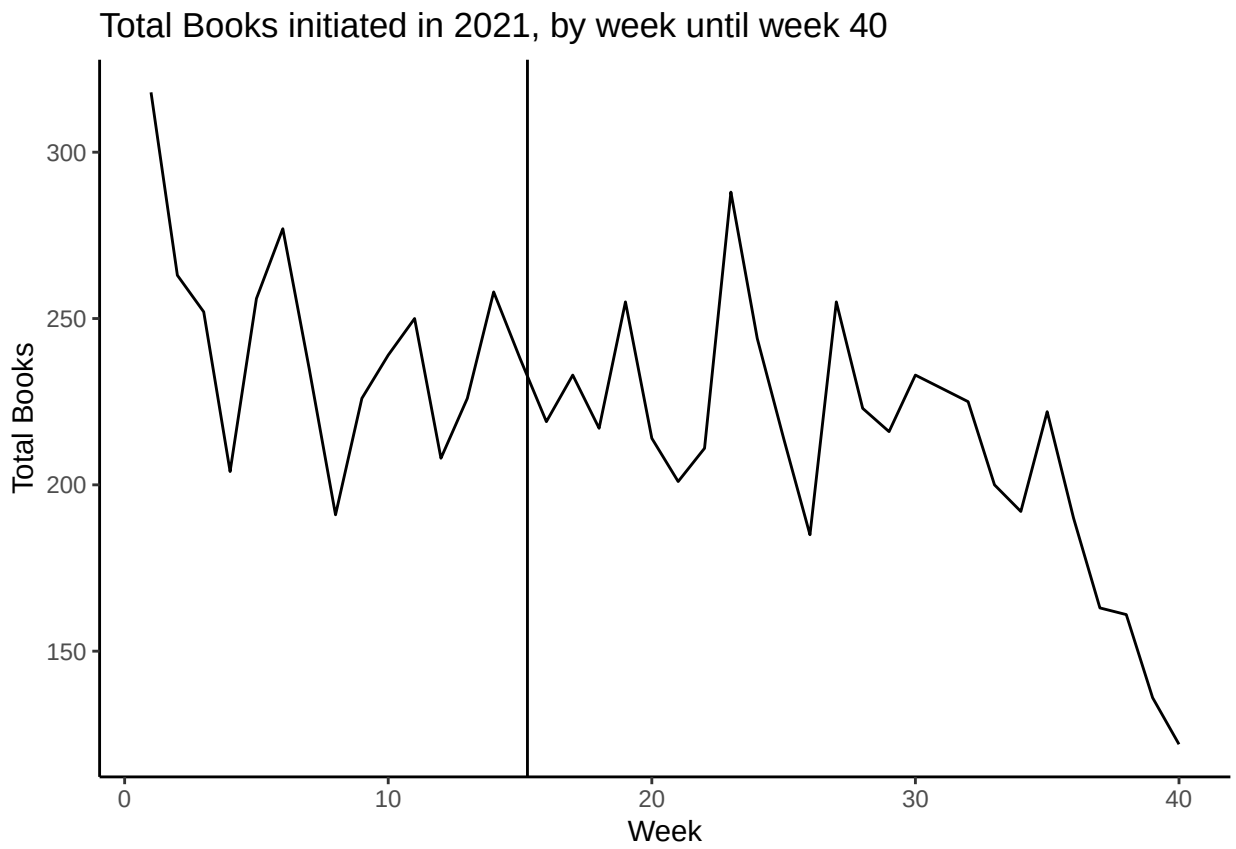
```
## [1] 4864
```

```r
books_per_reader2021_second <- unique_books2021_second / total_readers_2021_second
print(books_per_reader2021_second)
```

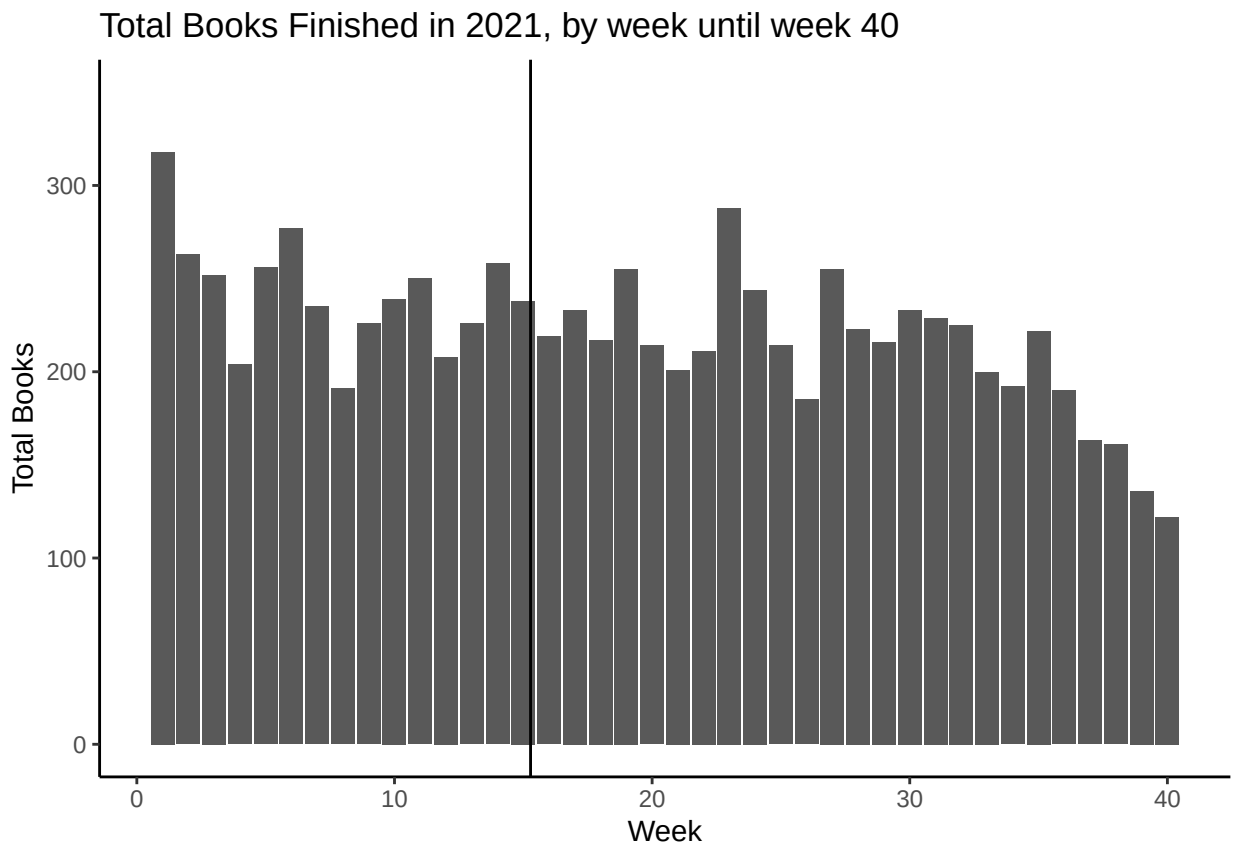
```
## [1] 17.81818
```

Represent the evolution of books started in 2021 in a line and bar plot

```r
plot1_2021_second <- goodreads_2021 %>%
  group_by(week_of_year) %>%
  select(week_of_year, book.url) %>%
  summarise(total_books = n())



graph_2021_line_second <- ggplot(plot1_2021_second, aes(x = week_of_year,
                                                        y = total_books)) +
  geom_line() +
  xlim(1, 40) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books initiated in 2021, by week until week 40")

graph_2021_line_second
```




```
graph_2021_bar_second <- ggplot(plot1_2021_second, aes(x = week_of_year,
                                                       y = total_books)) +
  geom_bar(stat = "identity", position = position_dodge(width = 1)) +
  ylim(0, 350) +
  geom_vline(xintercept = 15.282) +
  theme_bw() +
  theme(
    panel.border = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")
  ) +
  labs(y = "Total Books", x = "Week") +
  ggtitle("Total Books Finished in 2021, by week until week 40")

graph_2021_bar_second
```




```
print(total_books2019_second)
```

```
## [1] 9326
```

```
print(total_readers_2019_second)
```

```
## [1] 437
```

```
print(unique_books2019_second)
```

```
## [1] 7863
```

```
print(unique_author2019_second)
```

```
## [1] 4346
```

```
print("------------------------")
```

```
## [1] "------------------------"
```

```r
print(total_books2020_second)
```

```
## [1] 12429
```

```r
print(total_readers_2020_second)
```

```
## [1] 496
```

```r
print(unique_books2020_second)
```

```
## [1] 9871
```

```r
print(unique_author2020_second)
```

```
## [1] 5416
```

```r
print("------------------------")
```

```
## [1] "------------------------"
```

```r
print(total_books2021_second)
```

```
## [1] 10620
```

```r
print(total_readers_2021_second)
```

```
## [1] 473
```

```r
print(unique_books2021_second)
```

```
## [1] 8428
```

```r
print(unique_author2021_second)
```

```
## [1] 4864
```

B4.Conclusions

We conclude that from 2019 to 2020 there is an increase of all variables, either for the total number of active users (from 437 to 496), total books read (from 9.326 to 12.429) and an increase of distinctive author's (from 4.346 to 5.416) and books (from 7.863 to 9.871). Therefore we conclude that the lockdown period lead people to change their reading habits, since users started to read more ans simultaneously used their extra-free time to search for new author's (increase variety). Secondly, although the year of 2021 is incomplete all mentioned variables present higher values in comparison to 2019 (active users: from 437 to 473; total books read: from

9.326 to 10.620; unique authors: from 4.346 to 4.864; unique books: from 7.863 to 8.428).

Finally the usage of date.read should be used and not date.added since values look more reliable although further theoretical investigation should be conducted to make such conclusion. A second possibility is to bundle date.added with date.read values for the respective book. This way the only instances used for research are the ones with date added before date read. We also believe it would be interesting to add a 9th variable to our data set for the number of pages per book, to avoid possible situations of books with 900 pages that are read and finished in the same day. Although possible is highly unlikable. A future data set will be uploaded in the beggining of 2021 with all the information regarding 2021 and with the new variable "book.pages", indicating the number of pages for the respective book.