# Analyzing Netflix History


### Nanna Marie Steenholdt


### 10/29/2020

While this code should work and provides a little help in case of bugs, there is a helpful commentary underneath each chunk and a README on my github here: https://github.com/nsteenholdt/CultData_ExamProject (https://github.com/nsteenholdt/CultData_ExamProject) There is also a paper about this code, which includes cultural importance and some interpretations with examples.

```
# Calling necessary pacakges
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(tidyr)
library(lubridate)
```

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

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggplot2)
```

If you are unable to call these libraries, it might be because you do not have these packages installed. Try install.packages() to resolve this.

```
# READING DATA FROM CSV DOWNLOADED FROM NETFLIX ACCOUNT
mynetflix <- read.csv("NetflixViewingHistory.csv")

# Take a look at your new dataframe
str(mynetflix)
mynetflix

# Edit the dates so it uses dashes instead of slashes - this will make further analys
is easier
# IF your dates are in month-date-year format, change the dmy() to mdy() and re-run t
he code
mynetflix$Date <- dmy(mynetflix$Date)

# To see where your current working directory is
getwd()
```

If R cannot find the file, it might be because the file you want to use is not located in your working directory. Use the getwd() to see where this RMarkdown is located on your computer, and move your .csv file to this location either via your shell or manually - alternatively, use setwd() to the desired location for this Rmarkdown. If you plan to use the netflix data provided already, beware that the .csv file will not move as well, when you set your working directory. It might also be that your file has another name, should you have changed it at some point - check that it is the correct in line 24, where you use read.csv().

```
## WHICH SHOWS DID YOU BINGEWATCH?

# We seperate the Titke column from our original dataframe into three columns: "Titl
e", "Season" and "Episode_title"
mynetflix_series <- mynetflix %>%
  separate(col = Title, into = c("Title", "Season", "Episode_title"), sep = ': ')
```

```
## Warning: Expected 3 pieces. Additional pieces discarded in 241 rows [38, 77, 85,
## 129, 130, 270, 271, 279, 280, 298, 299, 325, 326, 400, 412, 418, 427, 449, 452,
## 453, ...].
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 167 rows [3, 9,
## 196, 297, 380, 499, 533, 536, 564, 733, 739, 753, 754, 761, 762, 763, 764, 765,
## 808, 822, ...].
```

```
# We remove rows that have empty cells in the "Episode" and "Season" columns, because
if they are empty, that means they are a movie, not a series.
mynetflix_series <- mynetflix_series[!is.na(mynetflix_series$Season),]
mynetflix_series <- mynetflix_series[!is.na(mynetflix_series$Episode_title),]

# We register the number of episodes watched of the same series on the same date
mynetflix_binges <- mynetflix_series %>%
  count(Title, Date)

# If we consider "binge-watching" 6 or more episodes per day, we can choose only the
 rows that has a count ("n") higher than 6.
mynetflix_binges <- mynetflix_binges[mynetflix_binges$n >= 6,]
# We are now left with all the times we've bingewatched something. The instances belo
w are an alphabetical list.
mynetflix_binges

# We order our binge-watches by date
mynetflix_binges <- mynetflix_binges[order(mynetflix_binges$Date),]
# Now the list shows below shows the binge-watches from oldest to most recent
mynetflix_binges

# Alternatively we can order our binge watches by number of episodes watched on the s
ame date
mynetflix_binges1 <- mynetflix_binges[order(mynetflix_binges$n),]
# Now we can see them ordered from least to most amount of episodes
mynetflix_binges1
```

Note: if you disagree with how many episodes is considered a bingewatch, you will want to change the number in line 56 re-run the code from line 52 and onward.

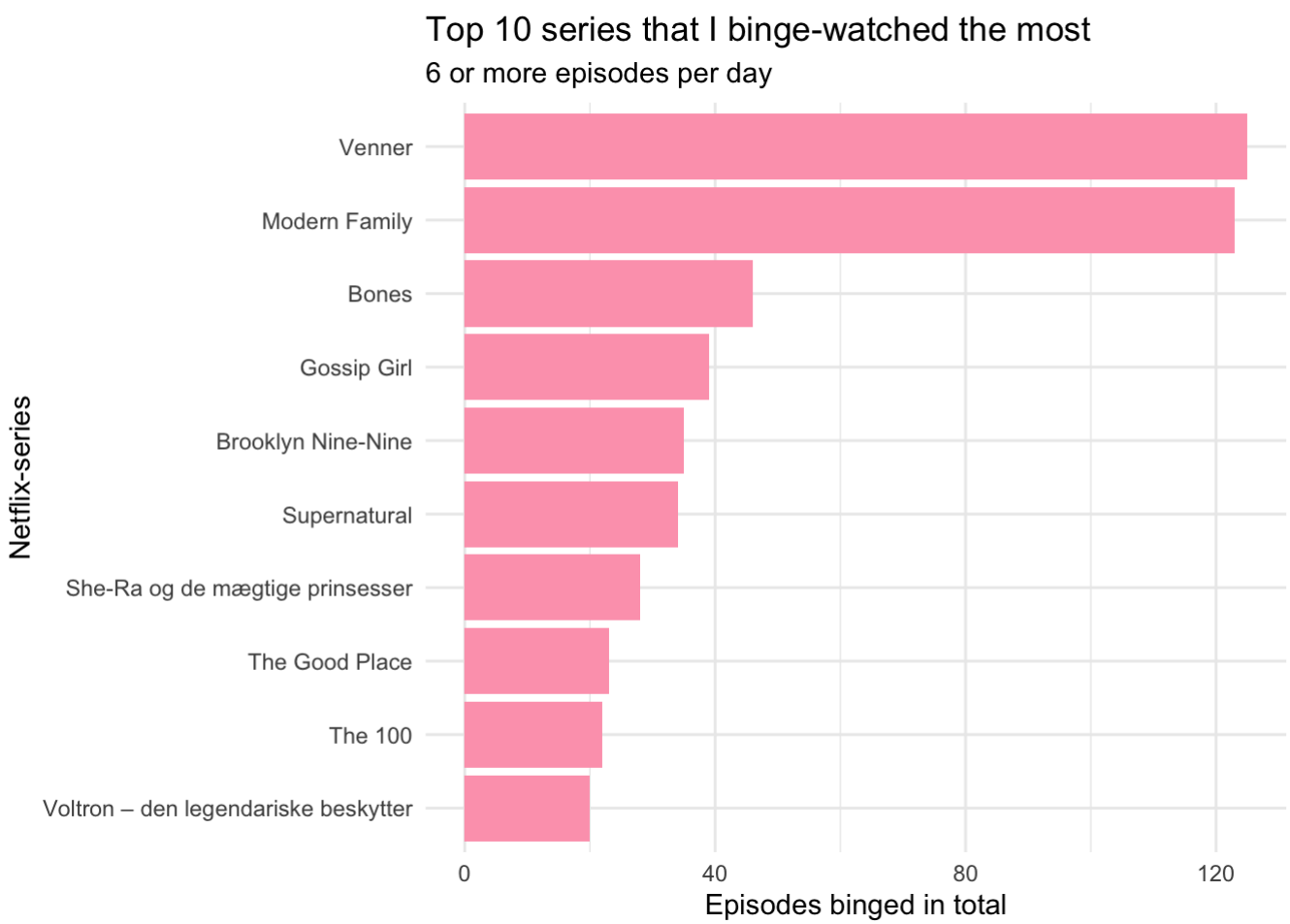
```
## VISUALIZING WHICH SERIES I BINGEWATCHED THE MOST

# I group the data by Title and sort by number of episodes watched
mynetflix_binges_all <- mynetflix_binges %>%
  group_by(Title) %>%
  summarise(n = sum(n)) %>%
  arrange(desc(n))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# I make a plot of my top 10 of series I binge-watched the most
mynetflix_binges_top <- mynetflix_binges_all %>%
  top_n(10) %>%
  ggplot(aes(x = reorder(Title, n), y = n)) +
  geom_col(fill = "#fda4ba") +
  coord_flip() + #This line switched the x and y coordinates in the plot
  ggtitle("Top 10 series that I binge-watched the most", "6 or more episodes per day"
) +
  labs(x = "Netflix-series", y = "Episodes binged in total") +
  theme_minimal()
```

```
## Selecting by n
```

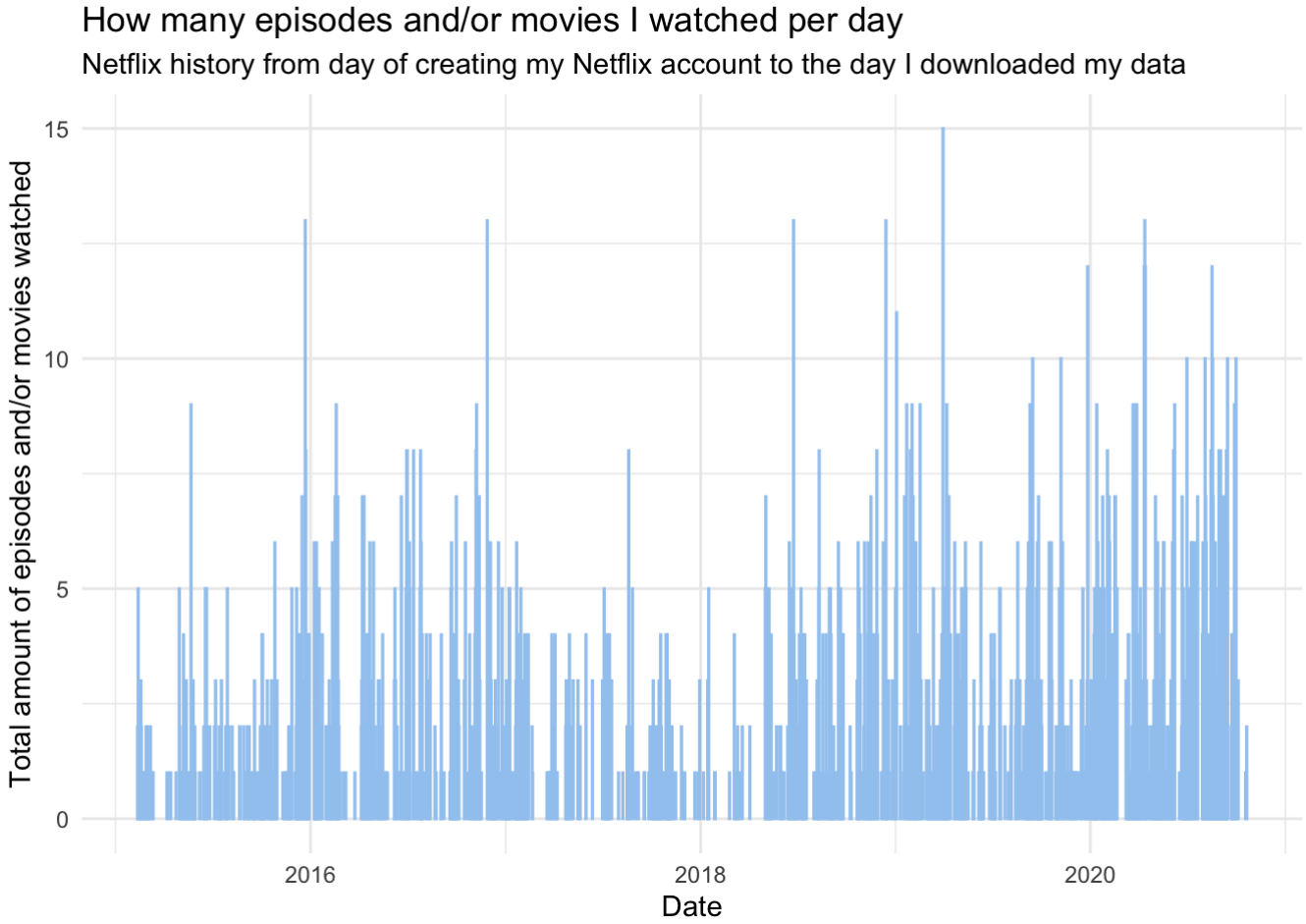
```
# Let's take a look at the plot
mynetflix_binges_top
```




Note: Not a fan of pink? Change the colour code in the ggplot geom_col() in line 87.

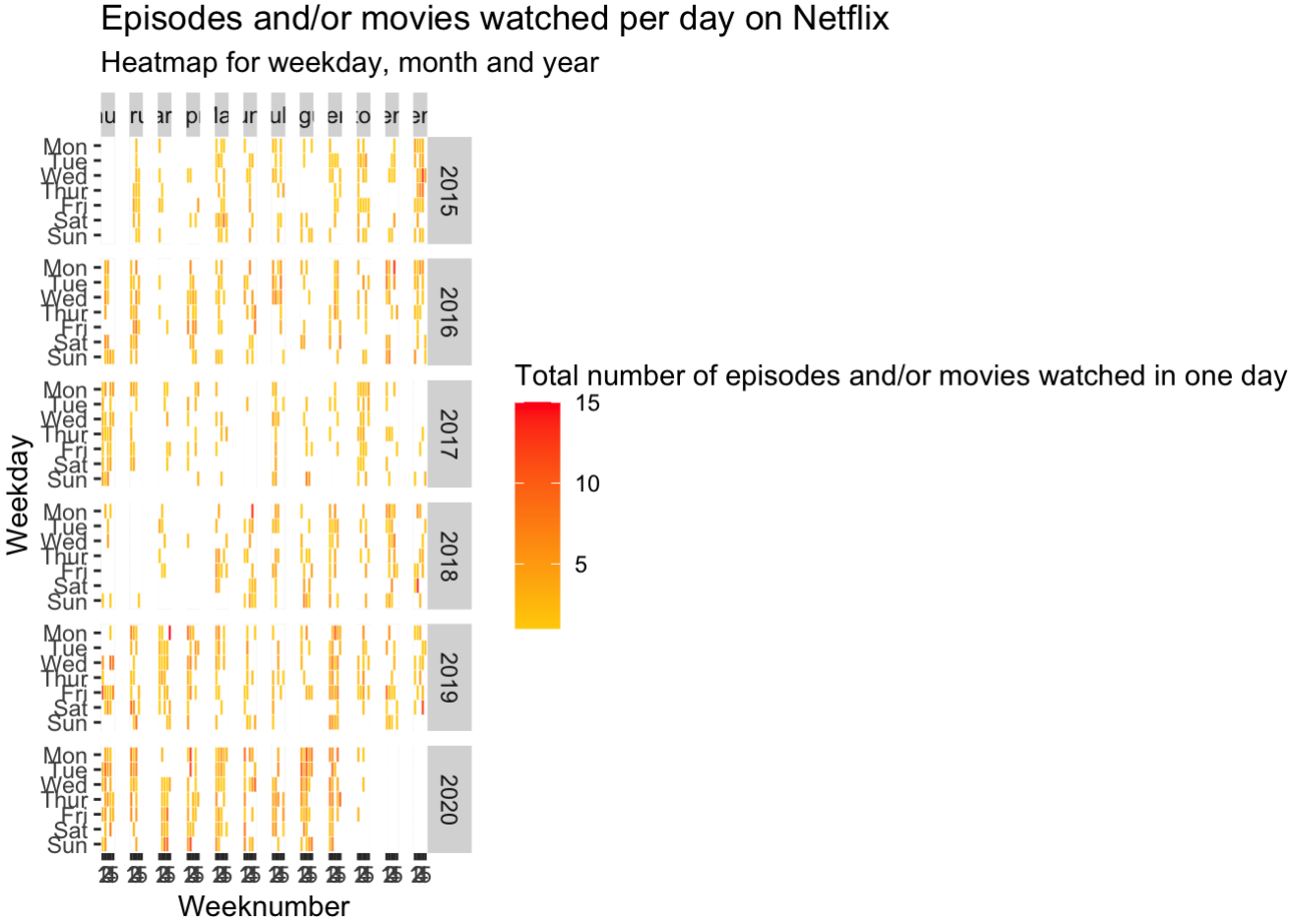
Note 2: It's important to notice that one might have watched more episodes than is counted in this plot. "Episodes binged in total" refers to the fact that this plot only includes the episodes that were part of a binge watch - so if you watched 1 episode of a show one day, it's not included in this plot.

```
# WE'VE SEEN WHICH SERIES WE BINGED THE MOST - BUT HOW MUCH NETFLIX DO YOU WATCH IN G
ENERAL

# I make a new dataframe that  counts the amount of episodes and/or movies watched pe
r day
netflix_episodes_day <- mynetflix %>%
  count(Date) %>%
  arrange(desc(n))

# I make a plot to visualize how many episodes and/or movies I watched per day
netflix_episodes_day_plot <- ggplot(aes(x = Date, y = n, color = n), data = netflix_e
pisodes_day) +
  geom_col(color = c("#a1caf1")) +
  theme_minimal() +
  ggtitle("How many episodes and/or movies I watched per day", "Netflix history from
 day of creating my Netflix account to the day I downloaded my data") +
  labs(x = "Date", y = "Total amount of episodes and/or movies watched")
# Let's take a look at the plot
netflix_episodes_day_plot
```




Note: The date range depends on how long you've had your Netflix account.

```
# CREATING A HEATMAP WHERE WE CAN SEE THE AMOUNT OF EPISODES AND/OR MOVIES WATCHES

# We order the Date column, so now it goes from oldest to newest dates.
netflix_episodes_day <- netflix_episodes_day[order(netflix_episodes_day$Date),]

# We make a column that has day of the week (Ranged 1 to 7)
netflix_episodes_day$day_week <- wday(netflix_episodes_day$Date)

# We add a column that has abbreviations of the day of the week (Mon to Sun)
netflix_episodes_day$day_weekF <- weekdays(netflix_episodes_day$Date, abbreviate = T)

# We add a column that has the month names (January to December)
netflix_episodes_day$monthF <- months(netflix_episodes_day$Date, abbreviate = T)

# We rename the columns to the abbreviations for each day. This step could be skipped
if you work in English but if you want to write them in another language, this line o
f code is useful.
netflix_episodes_day$day_weekF <-factor(netflix_episodes_day$day_week, levels = rev(1
:7), labels = rev(c("Mon","Tue","Wed","Thur","Fri","Sat","Sun")),ordered = TRUE)

# We rename the abbreviated months to the full names of the months. Once again, you c
an edit this according to languae.
netflix_episodes_day$monthF <- factor(month(netflix_episodes_day$Date),levels = as.ch
aracter(1:12), labels = c("January","February","March","April","May","June","July","A
ugust","September","October","November","December"),ordered = TRUE)

# We add a column that has the abbreviated month name and the year
netflix_episodes_day$year_month <- factor(as.yearmon(netflix_episodes_day$Date))

# We add a column that has the week number
netflix_episodes_day$week <- as.numeric(format(netflix_episodes_day$Date,"%W"))

# We add a column that divides the week numbers into which week number of a given wee
k it is
netflix_episodes_day$week_month <- ceiling(day(netflix_episodes_day$Date) / 7)

# We visualize the heatmap to see at which times we watched the most content
netflix_episodes_day_calendar <- ggplot(netflix_episodes_day, aes(week_month, day_wee
kF, fill = n)) +
  geom_tile(colour = "white") +
  facet_grid(year(netflix_episodes_day$Date) ~ monthF) +
  scale_fill_gradient(low = "#FFD000", high = "#FF1919") +
  ggtitle("Episodes and/or movies watched per day on Netflix", "Heatmap for weekday,
 month and year") +
  labs(x = "Weeknumber", y = "Weekday") +
  labs(fill = "Total number of episodes and/or movies watched in one day")
# Let's look at our heatmap
netflix_episodes_day_calendar
```



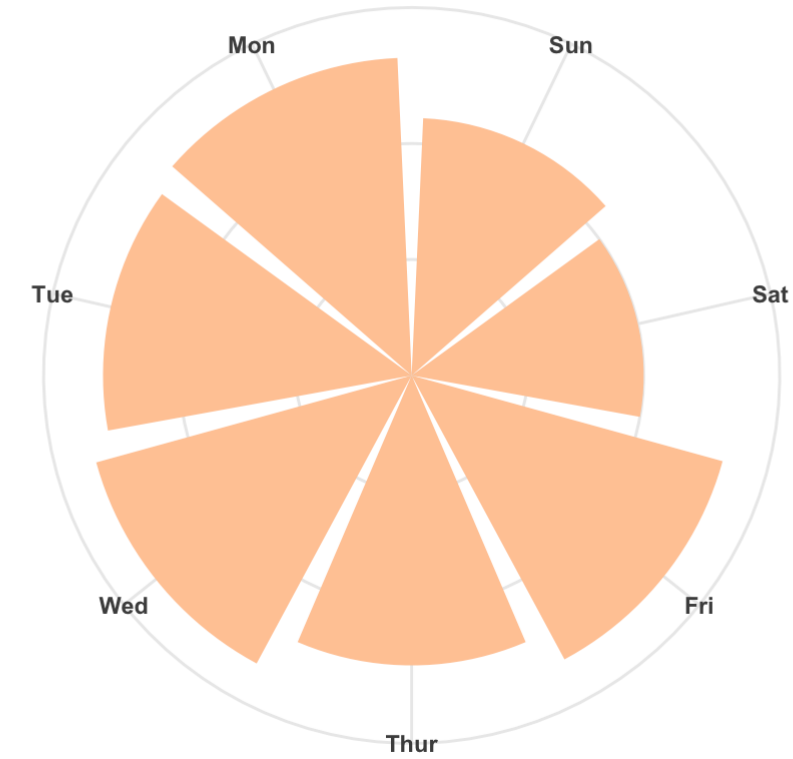

Note: confused about “weeknumber”? While usually a month is considered four weeks, some weeks overlap between months. Therefore there are 5 weeks in each month.

Note 2: If you knitted this file, this plot probably looks awful. I recommend rerunning the code in R and opening the plot manually for the full experience.

```
# ON WHICH WEEKDAY DID YOU WATCH THE MOST NETFLIX
# We make a dataframe that counts the amount of tiems you watched netflix on a partic
ular weekday
view_day <- netflix_episodes_day %>%
  count(day_weekF)
# We view the dataframe
view_day
```

```
##   day_weekF   n
## 1       Sun 111
## 2       Sat 100
## 3       Fri 139
## 4      Thur 125
## 5       Wed 141
## 6       Tue 133
## 7       Mon 137
```

```
# We visualize which weekday we watched netflix most frequently.
view_day_plot <- view_day %>%
  ggplot(aes(day_weekF, n)) +
  geom_col(fill = "#ffcba4") +
  coord_polar()  +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.text.x = element_text(face = "bold"),
        plot.title = element_text(size = 16, face = "bold")) +
  ggtitle("Frequency for weekday Netflix-viewings", "Activity per weekday on my Netfl
ix")
# We view the plot
view_day_plot
```




```
# DURING WHICH MONTH DID YOU WATCH NETFLIX MOST FREQUENTLY

# We make a dataframe that counts the amount of tiems you watched netflix during a pa
rticular month
view_month <- netflix_episodes_day %>%
  count(monthF)
# We view the dataframe
view_month
```

```
##       monthF  n
## 1    January 77
## 2   February 74
## 3      March 62
## 4      April 68
## 5        May 90
## 6       June 66
## 7       July 74
## 8     August 71
## 9  September 92
## 10   October 77
## 11  November 66
## 12  December 69
```

```
# We visualize which month we watched netflix most frequently.
view_month_plot <- view_month %>%
  ggplot(aes(monthF, n)) +
  geom_col(fill = "#A997DF") +
  coord_polar()  +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.text.x = element_text(face = "bold"),
        plot.title = element_text(size = 18, face = "bold")) +
  ggtitle("Frequency for monthly Netflix-viewings", "Activity per month on my Netfli
x")
# We view the plot
view_month_plot
```