Homework 8 Report

Neeraj Asthana (nasthan2)

All code is included in a file labelled “HW8.R”.

Reports are contained in “Homework 8 Report.pdf” and “Homework_8_Report.html” which were generated by “Homework 8 Report.Rmd”.

All graphics are included in the reports as well as in the Results folder.

```
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(ggplot2movies)
library(treemap)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

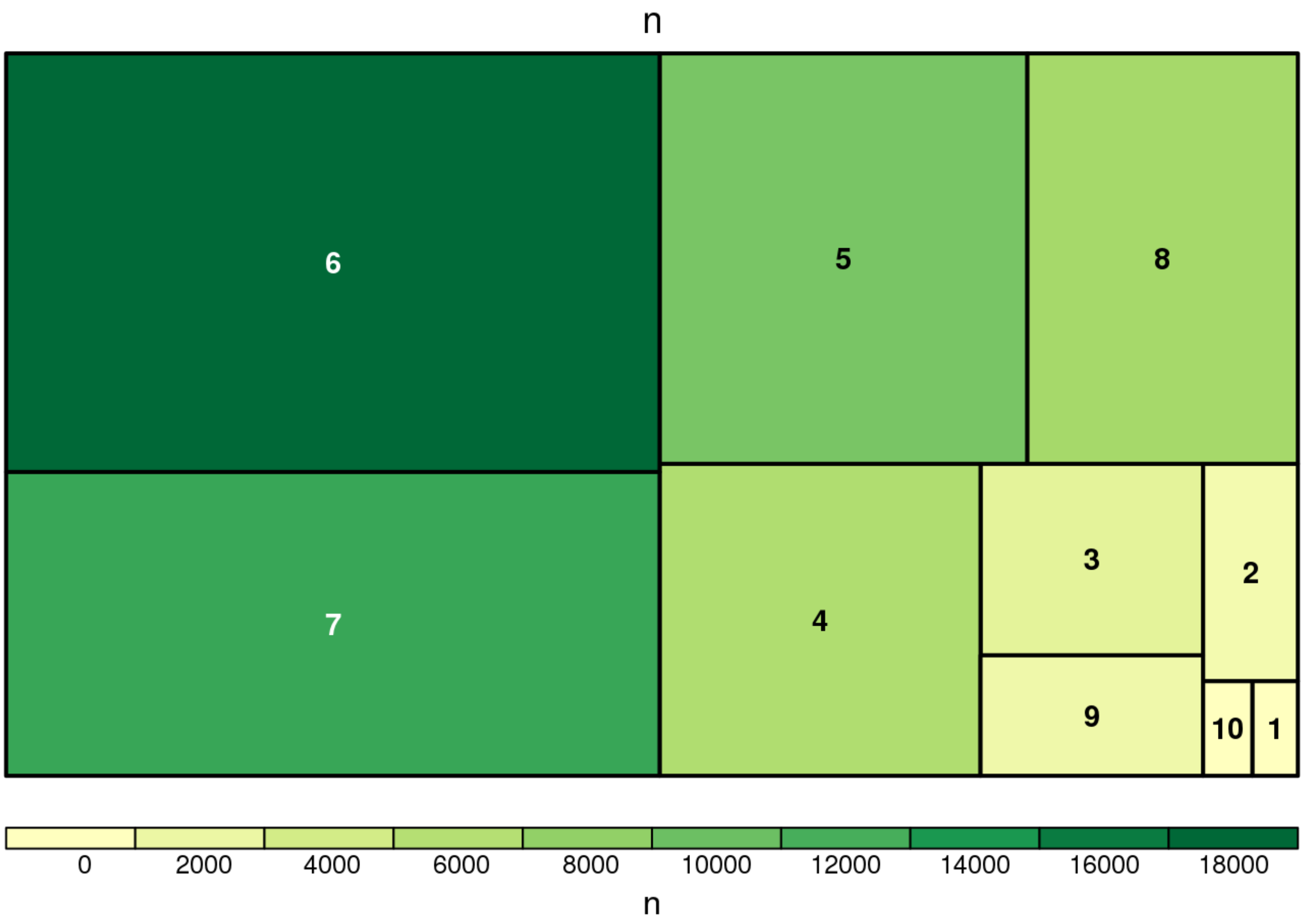
```
newdata<-ggplot2movies::movies
newdata["rating"]<-round(newdata["rating"])

newdata %>%
  group_by(rating, mpaa) %>%
  tally %>%
  ungroup -> mpaaratings
```

Exercise 1

Proportions of IMDB rating values:

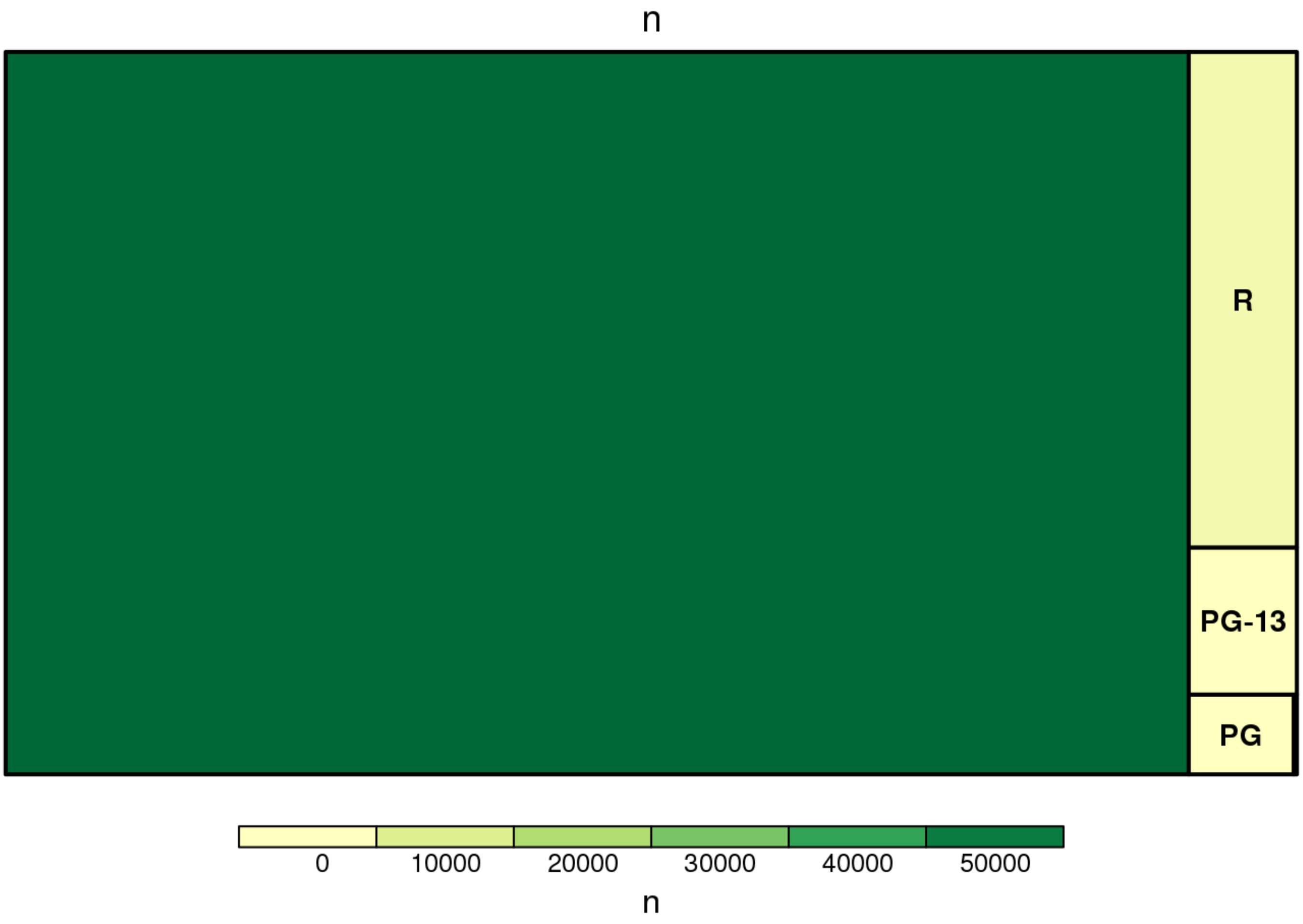
```
treemap(mpaaratings,
        index=c("rating"),
        vSize="n",
        vColor = "n",
        type="value")
```



Most of the movies in newdata have IMDB ratings of 6 and are therefore the most common. There is a very high relative proportion of movies that are rated 6 compared to the other ratings. The next largest rating group is 7. The next highest relative proportions of movies respectively are 5, 8, 4, 3, 9, 2, and 10. Lastly, the least relative proportion of movies have an IMDB rating of 1 and are therefore the least common. Therefore, in general most movies have an IMDB value between 4 and 8 as there is a relatively high proportion of these ratings. In general mot movies do not have an IMDB rating equal to or above a 9 or equal to or below a 3. Therefore, movies tend to not generally be very bad (1,2,3) or extremely good (9,10).

Proportions of MPAA rating values:

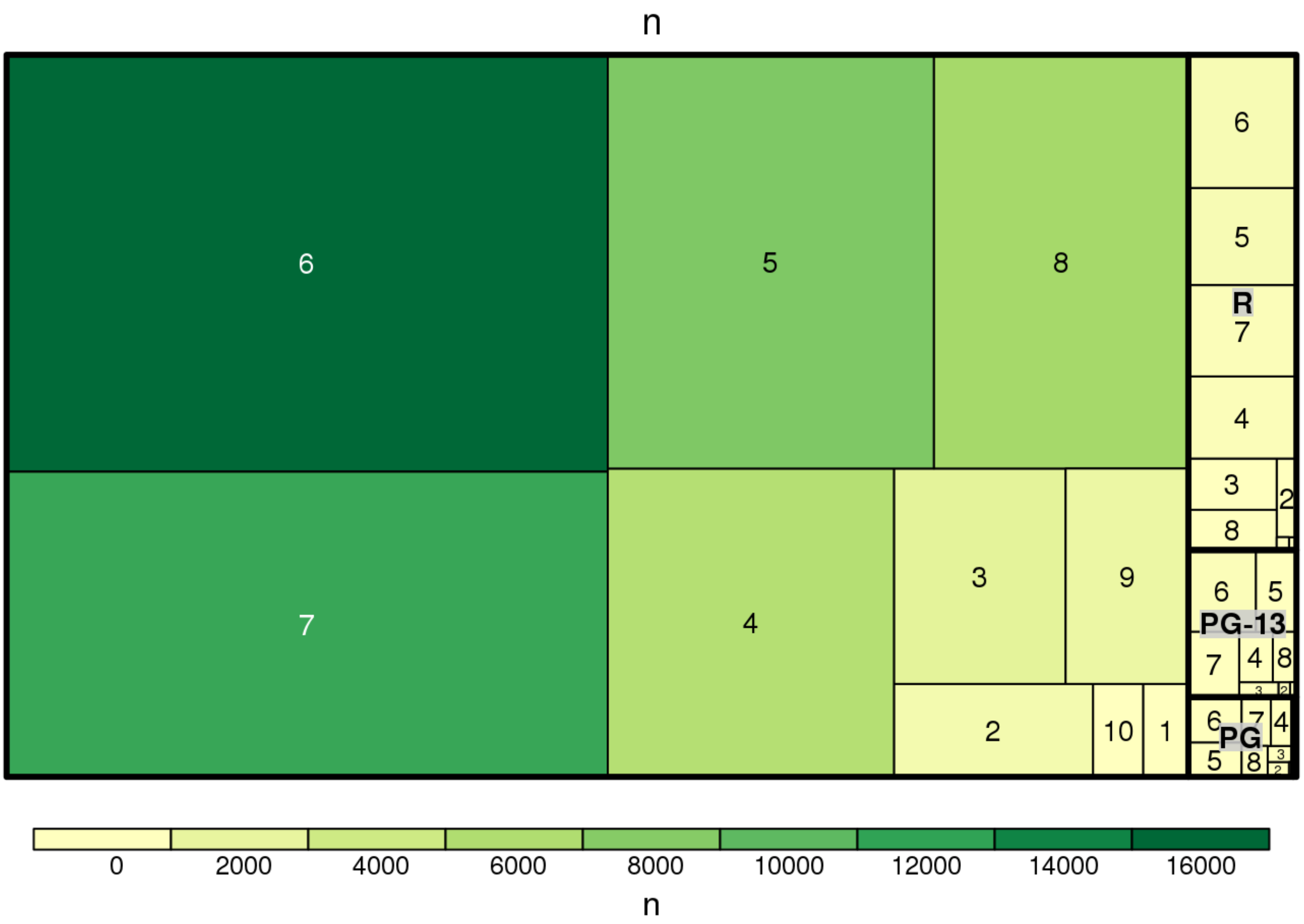
```
treemap(mpaaratings,
        index=c("mpaa"),
        vSize="n",
        vColor = "n",
        type="value")
```



Most of the movies in newdata have MPAA ratings that are uncategorized and are therefore the most common. There is a very high relative proportion of movies that are uncategorized compared to the other ratings. The next largest group is the “R” rated movies which have a much larger relative proportion that “PG-13”, “PG”, and, “NC-17”. “R” rated movies are followed by “PG-13” movies which are then followed by “PG” rated movies. Lastly, the least relative proportion of movies have an MPAA rating of “NC-17” and are therefore the least common.

Proportion of IMDB ratings within each MPAA rating group:

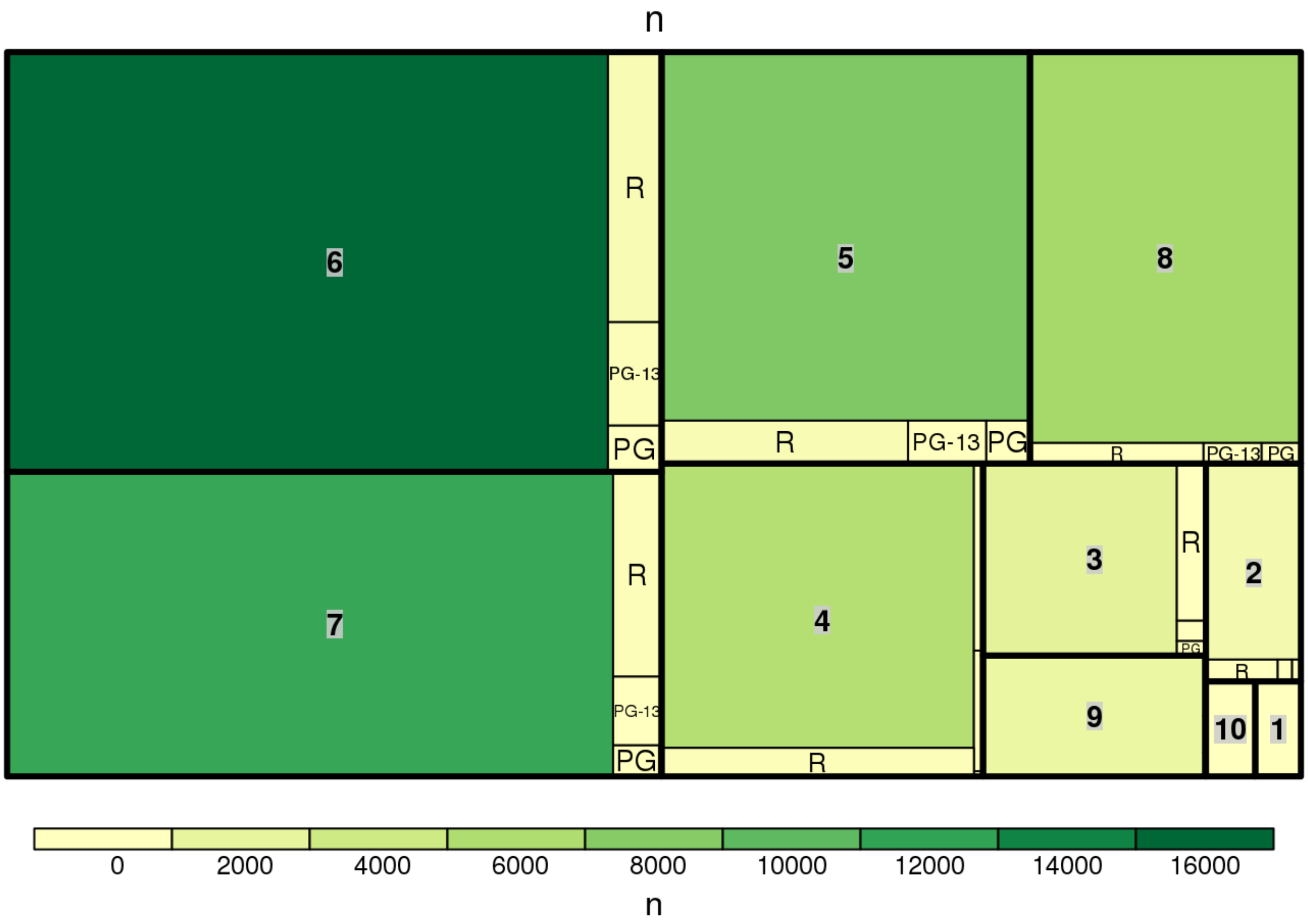
```
treemap(mpaaratings,
        index=c("mpaa", "rating"),
        vSize="n",
        vColor = "n",
        type="value")
```



Movies with an MPAA rating group of “Uncategorized” have the most proportion of IMDB ratings of 6 (which is the most common), followed by 7, 5, 4, 8, 3, 9, 2, and 10 respectively. The least for the “uncategorized” group is 1 (which is the least common. Movies with an MPAA rating group of “R” have the most proportion of IMDB ratings of 6 (which is the most common), followed by 5, 7, 4, 3, 8, and 2 respectively. The least for the “R” rated group is between 1, 9, and 10 which are not visible on the treemap (these are the least common). Movies with an MPAA rating group of “PG-13” have the most proportion of IMDB ratings of 6 (which is the most common), followed by 5, 7, 4, 8, 3, and 2 respectively. The least for the “PG-13” rated group is between 1, 9, and 10 which are not visible on the treemap (these are the least common). Movies with an MPAA rating group of “PG” have the most proportion of IMDB ratings of 6 (which is the most common), followed by 5, 7, 4, 8, 3, and 2 respectively. The least for the “PG” rated group is between 1, 9, and 10 which are not visible on the treemap (these are the least common). Movies with an MPAA rating group of “NC-17” are not even visible on the treemap. Most movies have an MPAA rating of “uncategorized” with IMDB ratings between 4 and 8. In general the IMDB ratings do not greatly differ among MPAA rating groups with the highest proportions of movies having an IMDB rating between 4 and 8 and lower proportions for the other rating values.

Proportion of MPAA rating group for each IMDB rating value:

```
treemap(mpaaratings,
        index=c("rating", "mpaa"),
        vSize="n",
        vColor = "n",
        type="value")
```



In general, proportions of MPAA ratings do not change across IMDB rating groups. The trend is similar across all IMDB rating groups as most of the groups have the largest proportion being uncategorized, the second largest being “R” rated movies, the third largest being “PG-13” rated movies, and the fourth largest being “PG” rated movies, and the smallest being “NC-17” rated movies. IMDB ratings of 1, 9, and 10 have very small proportions of ratings that are not uncategorized and therefore these other categories are not visible on the treemap.

Exercise 2

```
inst_pkgs = load_pkgs =  c("ggplot2","ggplot2movies", "dplyr","babynames","data.ta
ble","Rcpp")
git_pkgs = git_pkgs_load = c("streamgraph","DT")
load_pkgs = c(load_pkgs, git_pkgs_load)
pkgs_loaded = lapply(load_pkgs, require, character.only=T)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: babynames
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, last
```

```
## Loading required package: Rcpp
```

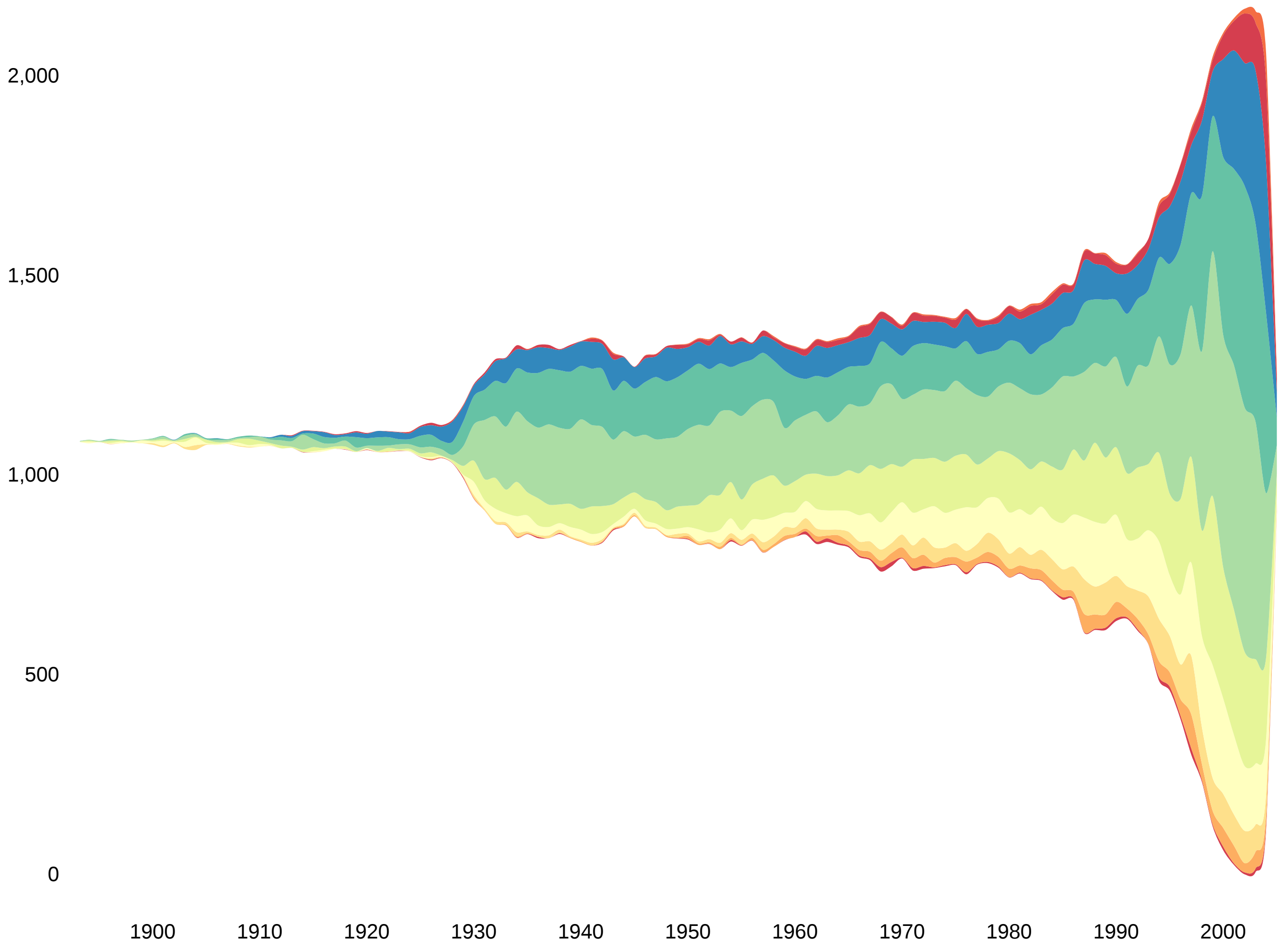
```
## Loading required package: streamgraph
```

```
## Loading required package: DT
```

Exercise 2a

```
newdata %>%
  group_by(year, rating) %>%
  tally() -> dat

streamgraph(dat, "rating", "n", "year") %>%
  sg_fill_brewer("Spectral") %>%
  sg_axis_x(tick_units = "year", tick_interval = 10, tick_format = "%Y")
```

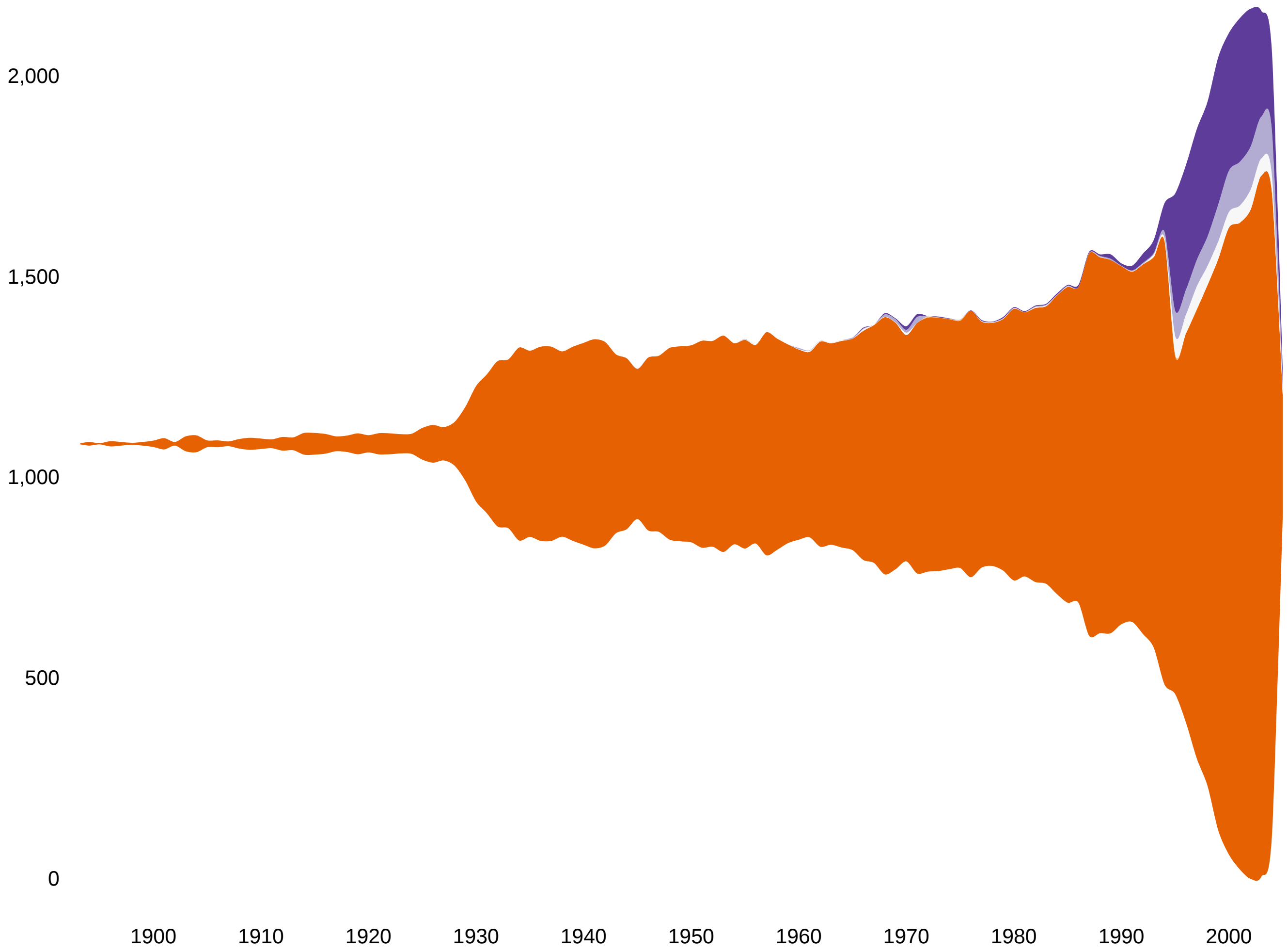


The plot is included in the Results folder under the label “Exercise2a.html”. In general we can see a dramatic increase in the number of movies created over the years. From 1893 to 1930, relatively few movies were made, however, there is a dramatic increase in the number of movies made and in 2003 a total of 2158 movies were made (largest). Over the years there appears to be large proportions of movies made that have an IMDB rating between a 3 and an 8 (around the same proportion each year with slight fluctuations but not significant). However, there appears to be relatively small proportions of movies that have an IMDB rating below 3 or above 10 (around the same proportion each year with slight fluctuations but not significant).

Exercise 2b

```
newdata %>%
  group_by(year, mpaa) %>%
  tally -> dat2

streamgraph(dat2, "mpaa", "n", "year") %>%
  sg_fill_brewer("PuOr") %>%
  sg_axis_x(tick_units = "year", tick_interval = 10, tick_format = "%Y")
```



The plot is included in the Results folder under the label “Exercise2b.html”. In general we can see a dramatic increase in the number of movies created over the years. From 1893 to 1930, relatively few movies were made, however, there is a dramatic increase in the number of movies made and in 2003 a total of 2158 movies were made (largest). Additionally, over the years we can see that there is a much higher proportion of “Uncategorized”" movies comparedd to other MPAA ratings (“R”, “PG-13”, “PG”, “NC-17”) for each of the years. It appears as though the MPAA ratings were not established till the 1960’s as all of the ratings other than “Uncategorized” grow in proportion after this time. There is also a dramatic dip in the number of “Uncategorized” movies in the early 1990’s (drop from 1105 to 842). There is also a dramatic increase in the number of “R” rated movies during this same time period (incrase from 71 to 293). It appears that during this time, MPAA ratings began being more enforced.

Exercise 3

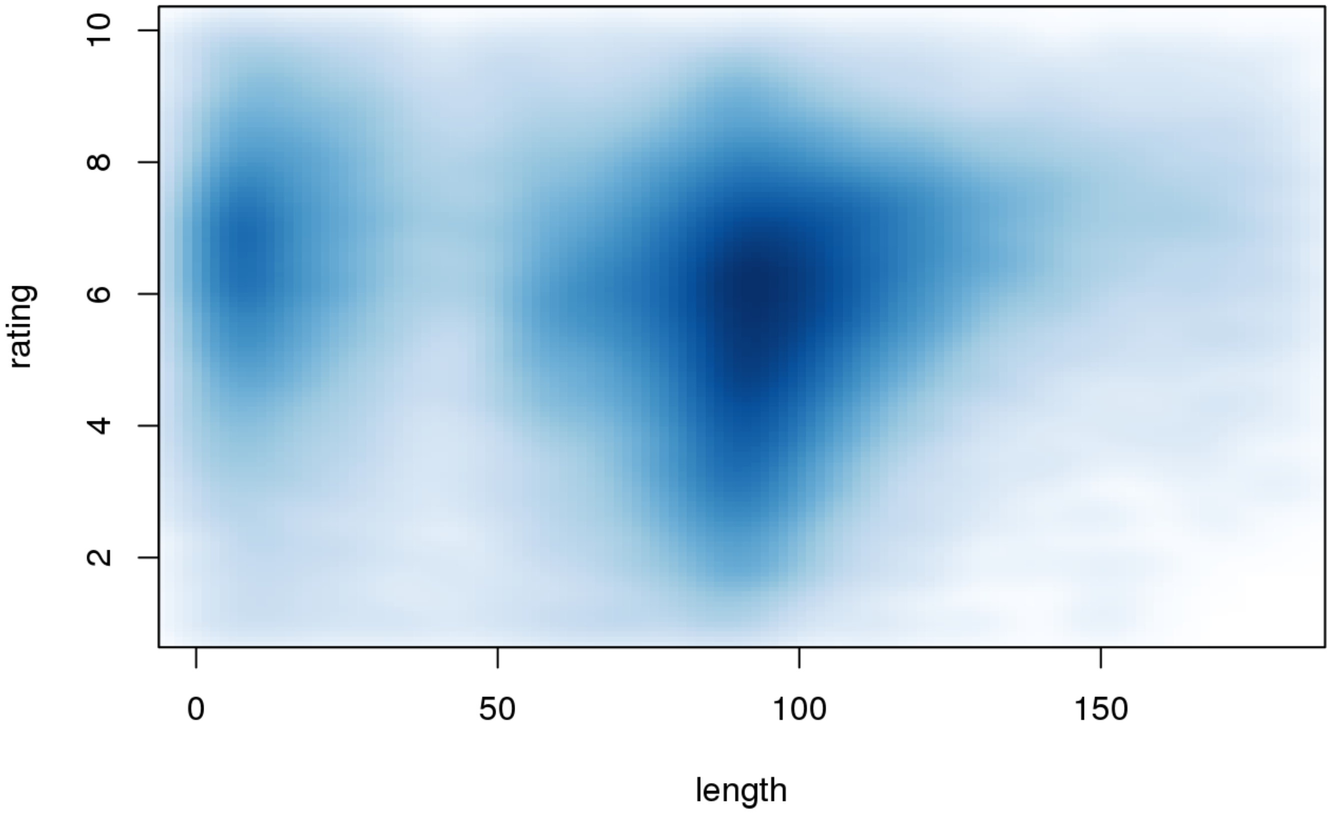
Exercise 3a

```
kdedata <- ggplot2movies::movies[,c("length", "rating")]
kdedata <- kdedata[kdedata[,"length"] <= 180,]
```

Created specificed “kdedata” dataset.

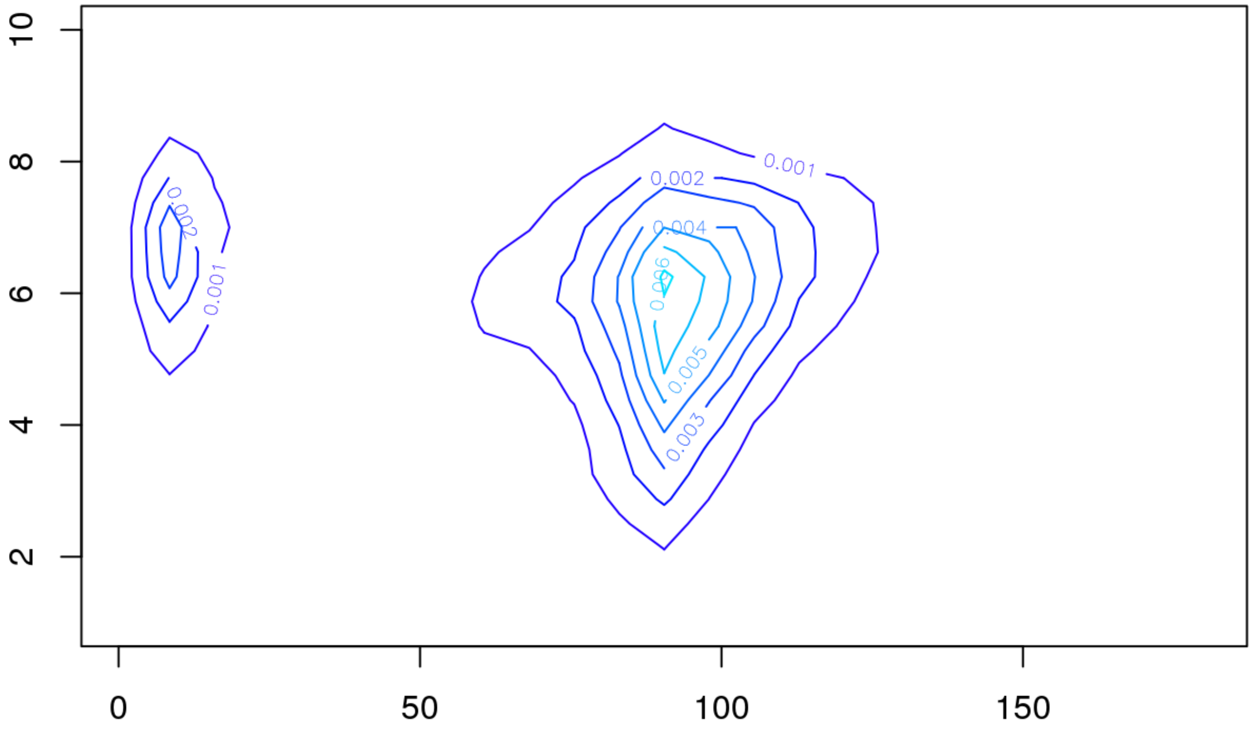
Exercise 3b

```
smoothScatter(kdedata[,c(1,2)], nrpoints=0)
```



```
band<-function(x)
{
  r <- quantile(x, c(0.25, 0.75))
  h <- (r[2] - r[1])/1.34
  4 * 1.06 * min(sqrt(var(x)), h) * length(x)^(-1/5)
}

h <- c(band(kdedata$length), band(kdedata$rating))
fit = kde2d(kdedata$length, kdedata$rating, h = h)
contour(fit, col = topo.colors(20))
```



Since the kernel density plot matches the smoothScatter well, I believe that my scales are well chosen. My scale values was 9.5136726 for the length and 0.7047165 for the rating (found using the band function in class). The density estimation tells us that most movies have a length between 80 and 1115 minutes and a rating a between 4 and 8. There is also a large density of movies that are 10 to 30 minutes in length and have a rating between 5 and 8. In general most movies seem to be 90 minutes in length and have a rating of 6.