Hw10

Cody

December 4, 2017

Here I consider the top rated movies of all time from IMDB.

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(purrr)
library(glue)
```

```
##
## Attaching package: 'glue'
```

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

```
library(stringr)

library(rvest)
```

```
## Warning: package 'rvest' was built under R version 3.4.2
```

```
## Loading required package: xml2
```

```
## Warning: package 'xml2' was built under R version 3.4.2
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:purrr':
##
```

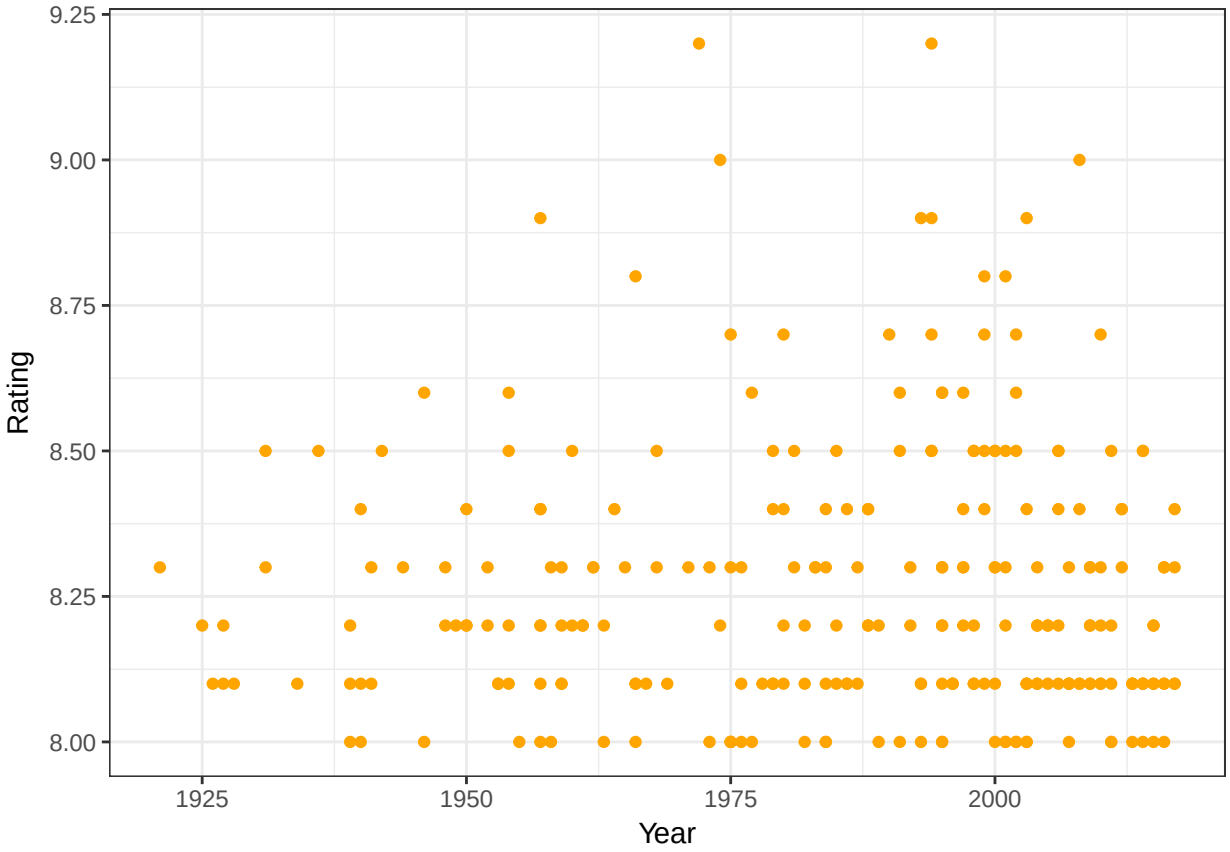
```
##     pluck
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(xml2)

url <- 'http://www.imdb.com/chart/top?ref_=nv_mv_250_6'
#Reading the HTML code from the website
webpage <- read_html(url)

rank_title <- webpage %>%
  html_nodes(".titleColumn") %>%
  html_text(trim = TRUE)
# Here I clean the artifacts of html
rank_title <- gsub("\n", "", rank_title)
rank_title <- gsub("      ", "", rank_title)
rank_title <- gsub("   ", "", rank_title)
rank_title <- gsub(".*\\.", "", rank_title)

# Extract year
year <- gsub("^.*\\(", "", rank_title)
year <- gsub(".{1}$", "", year) %>%
  as.numeric()

# Delete year off end
rank_title <- gsub(".{7}$", "", rank_title)

# Pull the ratings
rank_rating <- webpage %>%
  html_nodes(".ratingColumn.imdbRating") %>%
  html_text(trim = TRUE) %>%
  as.numeric()

# Put the clean data set together
rank_df <- data_frame(Title = rank_title,
                      Rating = rank_rating,
                      Year = year)

write.csv(rank_df, file = "movie_rankings.csv")
```

Now I have a dataset that contains the top 250 rated movies on IMDB as of 12/4/2017. This dataset has the movie title, the rating and the year all seperated into different columns.

Now to perform a super basic analysis,

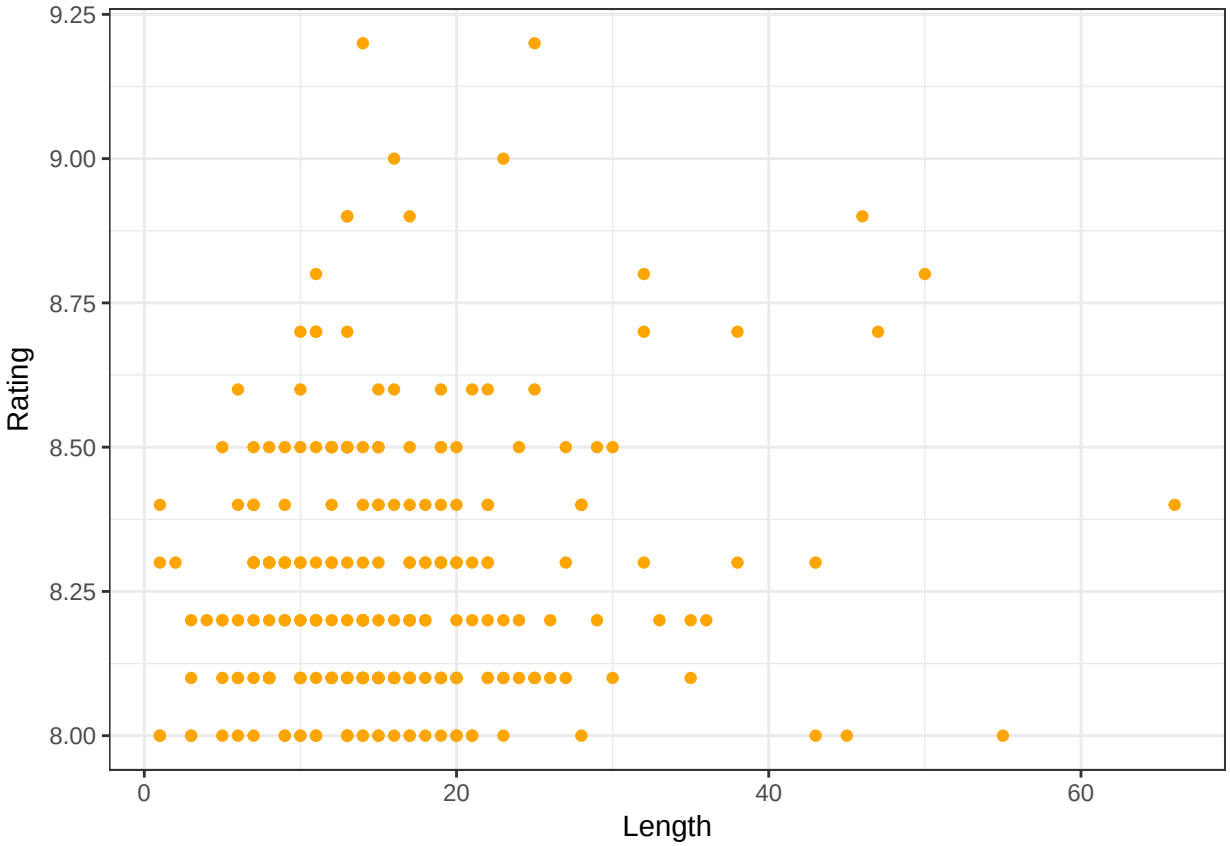
```
rank_df <- read.csv("movie_rankings.csv", as.is = TRUE)
rank_df$X <- NULL
ggplot(rank_df, aes(Year, Rating)) +
   geom_point(color="orange")+
    theme_bw()
```



From the scatterplot, we see quickly that there are only so many ratings that occur throughout our dataset and over time more movies appear on this list. We can possibly say from this that the quality of movies has increased over time, and therefore gets higher ratings. But again we are only looking at the top 250 rated of all time.

Now lets consider the length of the title as a potential influence on the data!

```
rank_df <- rank_df %>%
  mutate(Length = nchar(Title))
ggplot(rank_df, aes(Length, Rating)) +
   geom_point(color="orange")+
    theme_bw()
```



This seems not to be the case! Both the best and worst seem to have about 20 characters in their title. Longer titles don't seem to do much better.