

Data yang diambil dari website IMDb (Internet Movie Database, <https://www.imdb.com>) yang disusun berdasarkan popularitas film di tahun 2017 dengan user rating antara 5 dan 10 dan hanya diambil 100 film teratas. Data tersebut dapat diakses melalui link berikut:

https://www.imdb.com/search/title/?title_type=feature&release_date=2017-01-01,2017-12-31&user_rating=5.0,10.0&count=100

Langkah-langkah Web Scraping dengan R :

1. Install dan panggil package “xml2” dan “rvest” dengan perintah dalam R seperti berikut :

```
> install.packages("xml2")  
> library(xml2)  
> install.packages("rvest")  
> library(rvest)
```

Kemudian akan diperoleh hasil running sebagai berikut

```
Console Terminal Jobs x
~/
> install.packages("xml2")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/xml2_1.3.2.zip'
content type 'application/zip' length 3005803 bytes (2.9 MB)
downloaded 2.9 MB

package 'xml2' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
c:\Users\new\AppData\Local\Temp\RtmpaqIRMc\downloaded_packages
> library(xml2)
Warning message:
package 'xml2' was built under R version 4.0.5
> install.packages("rvest")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/rvest_1.0.0.zip'
content type 'application/zip' length 205989 bytes (201 KB)
downloaded 201 KB

package 'rvest' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
c:\Users\new\AppData\Local\Temp\RtmpaqIRMc\downloaded_packages
> library(rvest)
Warning message:
package 'rvest' was built under R version 4.0.5
> |
```

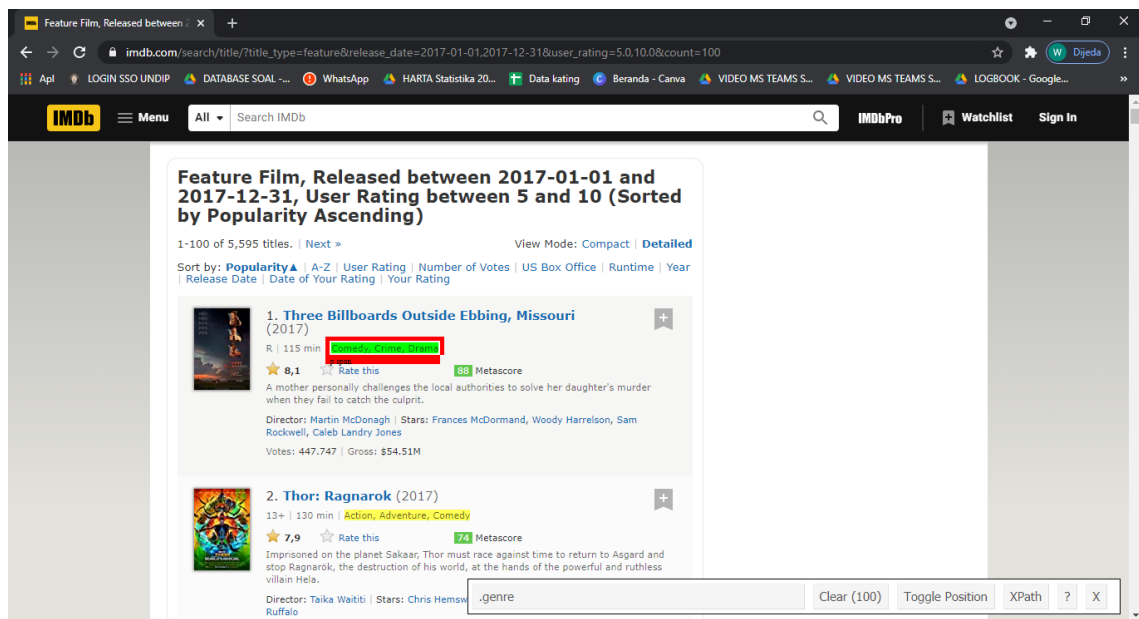
2. Kemudian lihat alamat dari website dan masukkan ke dalam R dengan perintah seperti berikut :

```
> alamat_website <-
"https://www.imdb.com/search/title/?title_type=feature&rele
ase_date=2017-01-01,2017-12-
31&user_rating=5.0,10.0&count=100"
> website <-read_html(alamat_website)
> website
```

Kemudian akan diperoleh hasil running sebagai berikut

```
Console Terminal Jobs x
~/
> alamat_website <- "https://www.imdb.com/search/title/?title_type=feature&release_date=2017-01-01,2017-12-31
&user_rating=5.0,10.0&count=100"
> website <-read_html(alamat_website)
> website
{html_document}
<html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fbml">
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">\n<script type="text ...
[2] <body id="styleguide-v2" class="fixed">\n          <img height="1" width="1" style="display:n ...
> |
```

3. Mengambil data genre yang ada pada halaman website dengan bantuan SelectorGadget



kemudian masukkan perintah ke dalam R seperti berikut :

```
> genre_data_website <-html_nodes(website, ".genre")
> genre_data_website
> data_genre <-html_text(genre_data_website)
> data_genre
> data_genre <-gsub("\n","",data_genre)
> data_genre
> data_genre <-gsub(" ","",data_genre)
> data_genre
> data_genre <-gsub(",.*","",data_genre)
> data_genre
> data_genre <-as.factor(data_genre)
> data_genre
```

Maka akan diperoleh hasil running sebagai berikut

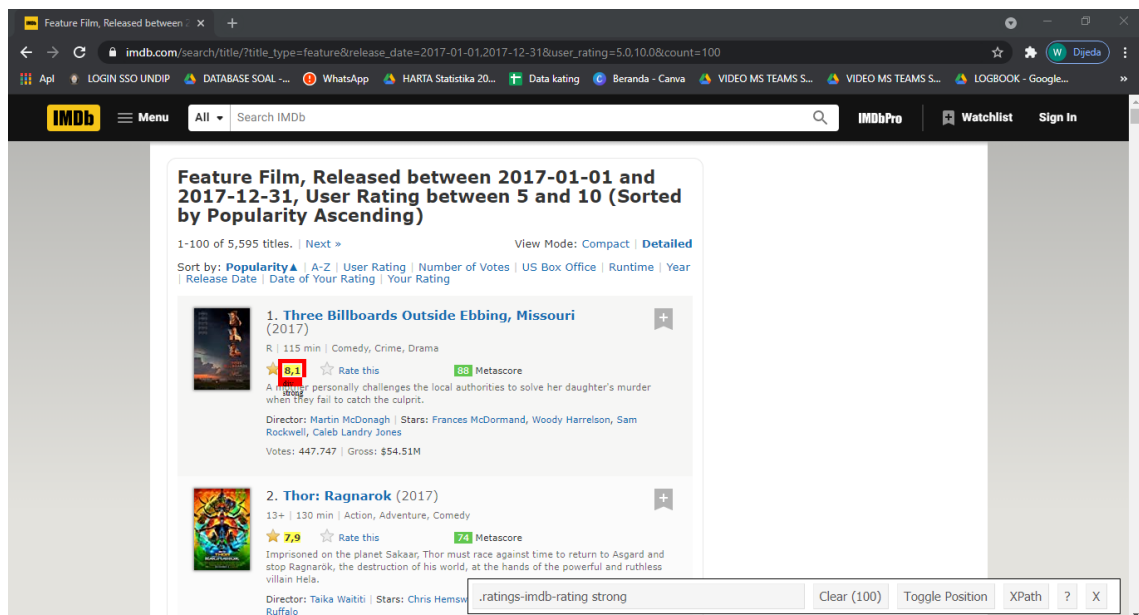
```
Console Terminal Jobs
~/R
> genre_data_website <-html_nodes(website, ".genre")
> genre_data_website
{xml_nodeset (100)}
[1] <span class="genre">\nComedy, Crime, Drama </span>
[2] <span class="genre">\nAction, Adventure, Comedy </span>
[3] <span class="genre">\nHorror, Mystery, Thriller </span>
[4] <span class="genre">\nAction, Drama, Mystery </span>
[5] <span class="genre">\nAction, Adventure, Fantasy </span>
[6] <span class="genre">\nAdventure, Drama, Fantasy </span>
[7] <span class="genre">\nDrama, Western </span>
[8] <span class="genre">\nAction, Adventure, Comedy </span>
[9] <span class="genre">\nAction, Adventure, Fantasy </span>
[10] <span class="genre">\nBiography, Drama, Musical </span>
[11] <span class="genre">\nAction, Adventure, Sci-Fi </span>
[12] <span class="genre">\nAction, Adventure, Comedy </span>
[13] <span class="genre">\nDrama, Romance </span>
[14] <span class="genre">\nAction, Crime, Drama </span>
[15] <span class="genre">\nHorror </span>
[16] <span class="genre">\nAction, Adventure, Crime </span>
[17] <span class="genre">\nAction, Adventure, Fantasy </span>
[18] <span class="genre">\nBiography, Drama, History </span>
[19] <span class="genre">\nHorror, Sci-Fi, Thriller </span>
[20] <span class="genre">\nAction, Adventure, Drama </span>
...
> data_genre <-html_text(genre_data_website)
> data_genre
[1] "\nComedy, Crime, Drama"
[3] "\nHorror, Mystery, Thriller"
[5] "\nAction, Adventure, Fantasy"
[7] "\nDrama, Western"
[9] "\nAction, Adventure, Fantasy"
[11] "\nAction, Adventure, Sci-Fi"
[13] "\nDrama, Romance"
[15] "\nHorror"
[17] "\nAction, Adventure, Fantasy"
[19] "\nHorror, Sci-Fi, Thriller"
[21] "\nAction, Thriller"
[23] "\nAction, Drama, History"
[25] "\nAction, Adventure, Comedy"
[27] "\nAction, Drama, Mystery"
[29] "\nAction, Adventure, Fantasy"
[31] "\nAction, Adventure, Comedy"
[33] "\nBiography, Drama, Musical"
[35] "\nAction, Adventure, Crime"
[37] "\nBiography, Drama, History"
[39] "\nAction, Adventure, Drama"
[41] "\nBiography, Comedy, Drama"
[43] "\nAnimation, Adventure, Comedy"
```

```

> data_genre <-gsub("\n","",data_genre)
> data_genre
[1] "Comedy, Crime, Drama" "Action, Adventure, Comedy" "Horror, Mystery, Thriller" "Action, Drama, Mystery"
[3] "Horror, Mystery, Thriller" "Action, Drama, Mystery" "Adventure, Drama, Fantasy" "Adventure, Drama, Fantasy"
[5] "Action, Adventure, Fantasy" "Action, Adventure, Comedy" "Action, Adventure, Comedy" "Action, Adventure, Comedy"
[7] "Drama, Western" "Biography, Drama, Musical" "Biography, Drama, Musical" "Biography, Drama, Musical"
[9] "Action, Adventure, Fantasy" "Action, Adventure, Comedy" "Action, Adventure, Comedy" "Action, Adventure, Comedy"
[11] "Action, Adventure, Sci-Fi" "Action, Adventure, Comedy" "Action, Adventure, Comedy" "Action, Adventure, Comedy"
[13] "Drama, Romance" "Action, Crime, Drama" "Action, Crime, Drama" "Action, Crime, Drama"
[15] "Horror" "Action, Adventure, Crime" "Action, Adventure, Crime" "Action, Adventure, Crime"
[17] "Action, Adventure, Fantasy" "Biography, Drama, History" "Biography, Drama, History" "Biography, Drama, History"
[19] "Horror, Sci-Fi, Thriller" "Action, Adventure, Drama" "Action, Adventure, Drama" "Action, Adventure, Drama"
[21] "Action, Thriller" "Biography, Comedy, Drama" "Biography, Comedy, Drama" "Biography, Comedy, Drama"
[23] "Action, Drama, History" "Animation, Adventure, Comedy" "Animation, Adventure, Comedy" "Animation, Adventure, Comedy"
[25] "Animation, Adventure, Family" "Action, Comedy, Crime" "Action, Comedy, Crime" "Action, Comedy, Crime"
[27] "Drama, Family" "Family, Fantasy, Musical" "Family, Fantasy, Musical" "Family, Fantasy, Musical"
[29] "Action, Adventure, Fantasy" "Action, Drama, Sci-Fi" "Action, Drama, Sci-Fi" "Action, Drama, Sci-Fi"
[31] "Action, Sci-Fi, Thriller" "Action, Adventure, Comedy" "Action, Adventure, Comedy" "Action, Adventure, Comedy"
[33] "Action, Adventure, Fantasy" "Action, Comedy, Crime" "Action, Comedy, Crime" "Action, Comedy, Crime"
[35] "Drama" "Action, Adventure, Sci-Fi" "Action, Adventure, Sci-Fi" "Action, Adventure, Sci-Fi"
[37] "Action, Adventure, Drama" "Comedy, Drama" "Comedy, Drama" "Comedy, Drama"
[39] "Crime, Drama, Mystery" "Comedy, Drama, History" "Comedy, Drama, History" "Comedy, Drama, History"
[41] "Crime, Drama, Mystery" "Action, Crime, Thriller" "Action, Crime, Thriller" "Action, Crime, Thriller"
[43] "Adventure, Biography, Crime" "Action, Adventure, Fantasy" "Action, Adventure, Fantasy" "Action, Adventure, Fantasy"
[45] "Drama, Romance" "Biography, Crime, Drama" "Biography, Crime, Drama" "Biography, Crime, Drama"
[47] "Drama, Horror, Mystery" "Comedy, Crime, Drama" "Comedy, Crime, Drama" "Comedy, Crime, Drama"
[49] "Horror, Mystery, Thriller" "Action, Adventure, Sci-Fi" "Action, Adventure, Sci-Fi" "Action, Adventure, Sci-Fi"
[51] "Crime, Drama, Thriller" "Biography, Drama, History" "Biography, Drama, History" "Biography, Drama, History"
[53] "Crime, Drama, Thriller" "Drama" "Drama" "Drama"
[55] "Drama, Horror, Mystery" "Horror, Sci-Fi, Thriller" "Horror, Sci-Fi, Thriller" "Horror, Sci-Fi, Thriller"
[57] "Action, Comedy, Crime" "Action, Thriller" "Action, Thriller" "Action, Thriller"
[59] "Action, Biography, Thriller" "Drama, Horror, Thriller" "Drama, Horror, Thriller" "Drama, Horror, Thriller"
[61] "Comedy, Horror" "Drama, Sci-Fi, Thriller" "Drama, Sci-Fi, Thriller" "Drama, Sci-Fi, Thriller"
[63] "Action, Adventure, Fantasy" "Action, Crime, Fantasy" "Action, Crime, Fantasy" "Action, Crime, Fantasy"
[65] "Action, Adventure, Drama" "Action, Crime, Drama" "Action, Crime, Drama" "Action, Crime, Drama"
[67] "Action, Sci-Fi, Thriller" "Action, Biography, Drama" "Action, Biography, Drama" "Action, Biography, Drama"
[69] "Horror, Mystery, Thriller" "Drama" "Drama" "Drama"
[71] "Horror, Mystery, Thriller" "Action, Fantasy, Thriller" "Action, Fantasy, Thriller" "Action, Fantasy, Thriller"
[73] "Biography, Comedy, Drama" "Drama, Romance" "Drama, Romance" "Drama, Romance"
[75] "Crime, Drama" "Adventure, Drama, Horror" "Adventure, Drama, Horror" "Adventure, Drama, Horror"
[77] "Crime, Horror, Mystery" "Drama, Western" "Drama, Western" "Drama, Western"
[79] "Comedy, Drama" "Action, Thriller" "Action, Thriller" "Action, Thriller"
[81] "Action, Adventure, Fantasy" "Comedy, Horror, Mystery" "Comedy, Horror, Mystery" "Comedy, Horror, Mystery"
[83] "Comedy, Music" "Action, Horror, Thriller" "Action, Horror, Thriller" "Action, Horror, Thriller"
[85] "Crime, Drama, Thriller" "Action, Adventure, Biography" "Action, Adventure, Biography" "Action, Adventure, Biography"
[87] "Drama, Thriller, War" "Drama, Romance" "Drama, Romance" "Drama, Romance"
[89] "Comedy, Drama" "Animation, Adventure, Comedy" "Animation, Adventure, Comedy" "Animation, Adventure, Comedy"
[91] "Action, Comedy, Crime" "Drama, War" "Drama, War" "Drama, War"
[93] "Comedy, Drama, Romance" "Comedy, Drama" "Comedy, Drama" "Comedy, Drama"
[95] "Biography, Crime, Drama" "Action, Adventure, Drama" "Action, Adventure, Drama" "Action, Adventure, Drama"
[97] "Drama, Romance" "Drama, History, Thriller" "Drama, History, Thriller" "Drama, History, Thriller"
[99] "Crime, Drama, History" "Animation, Action, Comedy" "Animation, Action, Comedy" "Animation, Action, Comedy"
> data_genre <-gsub(" ", "", data_genre)
> data_genre
[1] "Comedy,Crime,Drama" "Action,Adventure,Comedy" "Horror,Mystery,Thriller" "Action,Drama,Mystery"
[3] "Action,Adventure,Fantasy" "Adventure,Drama,Fantasy" "Drama,Western" "Action,Adventure,Comedy"
[5] "Action,Adventure,Fantasy" "Action,Crime,Drama" "Horror" "Action,Adventure,Comedy"
[7] "Action,Adventure,Fantasy" "Biography,Drama,History" "Horror,Sci-Fi,Thriller" "Action,Adventure,Drama"
[9] "Action,Thriller" "Biography,Comedy,Drama" "Action,Drama,History" "Animation,Adventure,Comedy"
[11] "Animation,Adventure,Family" "Action,Comedy,Crime" "Drama,Family" "Family,Fantasy,Musical"
[13] "Action,Adventure,Fantasy" "Action,Adventure,Sci-Fi" "Drama,Sci-Fi,Thriller" "Action,Adventure,Comedy"
[15] "Action,Adventure,Fantasy" "Action,Comedy,Crime" "Drama" "Action,Adventure,Sci-Fi"
[17] "Action,Adventure,Drama" "Comedy,Drama" "Crime,Drama,Mystery" "Comedy,Drama,History"
[19] "Crime,Drama,Mystery" "Action,Crime,Thriller" "Adventure,Biography,Crime" "Action,Adventure,Fantasy"
[21] "Drama,Romance" "Biography,Crime,Drama" "Drama,Horror,Mystery" "Comedy,Crime,Drama"
[23] "Horror,Mystery,Thriller" "Action,Adventure,Sci-Fi" "Crime,Drama,Thriller" "Biography,Drama,History"
[25] "Crime,Drama,Thriller" "Drama" "Drama,Horror,Mystery" "Horror,Sci-Fi,Thriller"
[27] "Action,Comedy,Crime" "Action,Thriller" "Action,Biography,Thriller" "Drama,Horror,Thriller"
[29] "Comedy,Horror" "Drama,Sci-Fi,Thriller" "Action,Adventure,Fantasy" "Action,Crime,Fantasy"
[31] "Action,Adventure,Drama" "Action,Crime,Drama" "Action,Sci-Fi,Thriller" "Action,Biography,Drama"
[33] "Horror,Mystery,Thriller" "Drama" "Horror,Mystery,Thriller" "Action,Fantasy,Thriller"
[35] "Biography,Comedy,Drama" "Drama,Romance" "Crime,Drama" "Adventure,Drama,Horror"
[37] "Crime,Horror,Mystery" "Drama,Western" "Comedy,Drama" "Action,Thriller"
[39] "Action,Adventure,Fantasy" "Comedy,Horror,Mystery" "Comedy,Music" "Action,Horror,Thriller"
[41] "Crime,Drama,Thriller" "Action,Adventure,Biography" "Drama,Thriller,War" "Drama,Romance"
[43] "Comedy,Drama" "Animation,Adventure,Comedy" "Action,Comedy,Crime" "Drama,War"
[45] "Comedy,Drama,Romance" "Comedy,Drama" "Biography,Crime,Drama" "Action,Adventure,Drama"
[47] "Drama,Romance" "Drama,History,Thriller" "Crime,Drama,History" "Animation,Action,Comedy"
> data_genre <-gsub(".", "", data_genre)
> data_genre
[1] "Comedy" "Action" "Horror" "Action" "Adventure" "Drama" "Action" "Action"
[3] "Biography" "Action" "Action" "Drama" "Action" "Horror" "Action" "Biography"
[5] "Horror" "Action" "Action" "Biography" "Action" "Animation" "Action" "Drama"
[7] "Family" "Action" "Action" "Action" "Action" "Action" "Action" "Drama"
[9] "Action" "Comedy" "Crime" "Comedy" "Crime" "Action" "Adventure" "Action"
[11] "Biography" "Drama" "Comedy" "Horror" "Action" "Crime" "Biography" "Crime"
[13] "Drama" "Horror" "Action" "Action" "Drama" "Comedy" "Drama" "Action"
[15] "Action" "Action" "Action" "Action" "Action" "Horror" "Drama" "Action"
[17] "Biography" "Drama" "Crime" "Adventure" "Crime" "Drama" "Comedy" "Action"
[19] "Comedy" "Comedy" "Action" "Crime" "Action" "Drama" "Drama" "Comedy"
[21] "Action" "Drama" "Comedy" "Comedy" "Biography" "Action" "Drama" "Crime"
[23] "Animation"
> data_genre <-as.factor(data_genre)
> data_genre
[1] Comedy Action Horror Action Action Adventure Drama Action Action Biography Action
[2] Action Drama Action Action Action Action Biography Horror Action Biography
[3] Action Animation Animation Action Drama Family Action Action Action Action
[4] Action Drama Action Action Action Comedy Crime Comedy Crime Action Adventure
[5] Drama Biography Drama Comedy Horror Action Crime Biography Crime Drama Drama
[6] Horror Action Action Action Drama Comedy Drama Biography Action Action Action
[7] Action Action Horror Drama Action Action Biography Drama Crime Adventure Crime
[8] Drama Comedy Action Action Comedy Comedy Action Crime Action Drama Drama
[9] Comedy Animation Action Drama Comedy Comedy Biography Action Drama Crime
[10] Animation
Levels: Action Adventure Animation Biography Comedy Crime Drama Family Horror

```

- Selanjutnya mengambil data rating yang ada pada halaman website dengan bantuan SelectorGadget



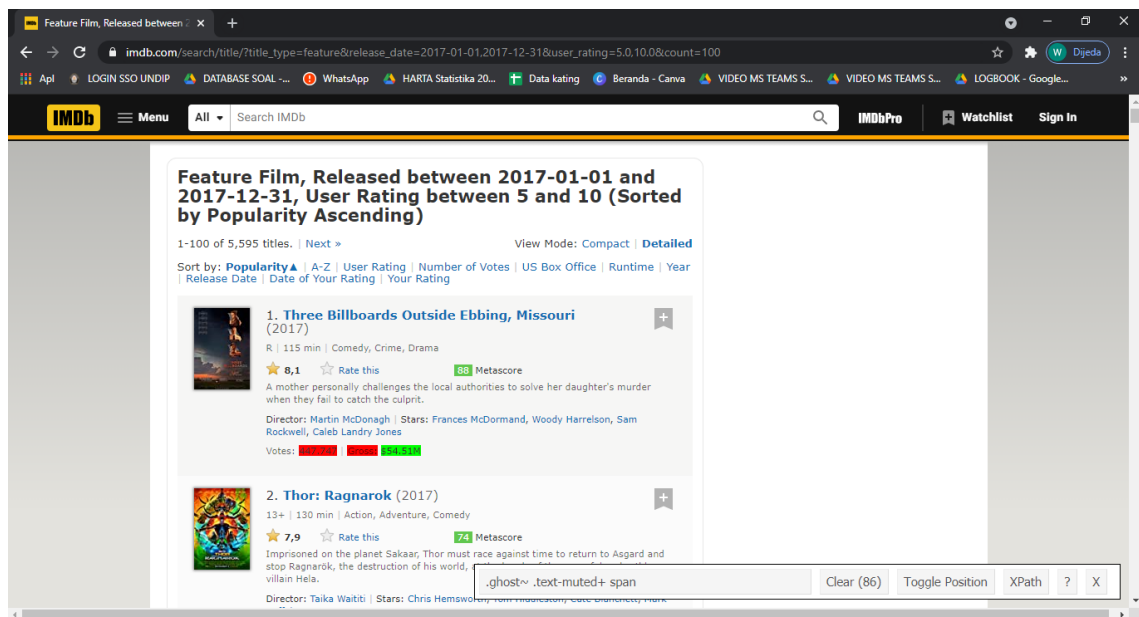
kemudian masukkan perintah ke dalam R seperti berikut :

```
> rating_data_website <-html_nodes(website, ".ratings-imdb-rating strong")
> rating_data_website
> data_rating <-html_text(rating_data_website)
> data_rating
> data_rating<- as.numeric(data_rating)
> data_rating
```

Dan akan diperoleh hasil running sebagai berikut

```
Console Terminal Jobs
~/R
> rating_data_website <-html_nodes(website, ".ratings-imdb-rating strong")
> rating_data_website
{xml_nodeset (100)}
[1] <strong>8.1</strong>
[2] <strong>7.9</strong>
[3] <strong>7.7</strong>
[4] <strong>8.0</strong>
[5] <strong>6.2</strong>
[6] <strong>7.3</strong>
[7] <strong>7.4</strong>
[8] <strong>6.9</strong>
[9] <strong>6.6</strong>
[10] <strong>7.6</strong>
[11] <strong>7.4</strong>
[12] <strong>7.6</strong>
[13] <strong>7.9</strong>
[14] <strong>7.6</strong>
[15] <strong>7.3</strong>
[16] <strong>6.6</strong>
[17] <strong>7.4</strong>
[18] <strong>7.0</strong>
[19] <strong>6.4</strong>
[20] <strong>6.7</strong>
...
> data_rating <-html_text(rating_data_website)
> data_rating
[1] "8.1" "7.9" "7.7" "8.0" "6.2" "7.3" "7.4" "6.9" "6.6" "7.6" "7.4" "7.6" "7.9" "7.6" "7.3" "6.6"
[17] "7.4" "7.0" "6.4" "6.2" "7.5" "7.8" "7.8" "8.4" "5.5" "8.0" "7.1" "6.5" "8.1" "5.3" "6.7"
[33] "6.5" "6.9" "7.6" "5.9" "7.4" "7.4" "6.5" "7.2" "7.7" "7.5" "7.2" "7.0" "7.5" "7.4" "6.6" "7.0"
[49] "6.3" "5.2" "7.4" "7.3" "7.2" "7.0" "6.6" "6.0" "6.7" "6.4" "6.5" "6.3" "5.4" "5.4" "6.9" "7.3" "6.3" "5.8" "7.6" "6.5" "7.6" "7.4" "6.3"
[65] "7.3" "6.3" "5.8" "7.6" "6.5" "7.6" "7.4" "6.3" "7.4" "6.6" "6.8" "5.7" "5.8" "7.2" "6.6" "7.0"
[81] "5.6" "6.6" "5.8" "6.4" "5.1" "6.7" "6.3" "6.3" "7.0" "6.7" "7.2" "7.4" "7.5" "7.2" "6.8" "6.4"
[97] "5.2" "7.4" "7.3" "7.3"
...
> data_rating<- as.numeric(data_rating)
> data_rating
[1] 8.1 7.9 7.7 8.0 6.2 7.3 7.4 6.9 6.6 7.6 7.4 7.6 7.9 7.6 7.3 6.6 7.4 7.0 6.4 6.7 6.2 7.5 7.8 7.8
[25] 8.4 5.5 8.0 7.1 6.5 8.1 5.3 6.7 6.5 6.9 7.6 5.9 7.4 7.4 6.5 7.2 7.7 7.5 7.2 7.0 7.5 7.4 6.6 7.0
[49] 6.3 5.2 7.4 7.3 7.2 7.0 6.6 6.0 6.7 6.4 6.5 6.3 5.4 5.4 6.9 7.3 6.3 5.8 7.6 6.5 7.6 7.4 6.3
[73] 7.4 6.6 6.8 5.7 5.8 7.2 6.6 7.0 5.6 6.6 5.8 6.4 5.1 6.7 6.3 6.3 7.0 6.7 7.2 7.4 7.5 7.2 6.8 6.4
[97] 5.2 7.4 7.3 7.3
```

5. Berikutnya mengambil data gross pendapatan pada halaman website dengan bantuan SelectorGadget



kemudian masukkan perintah ke dalam R seperti berikut :

```
> gross_data_website<- html_nodes(website, ".ghost~ .text-
muted+ span")
> gross_data_website
> data_gross<- html_text(gross_data_website)
> data_gross
> data_gross<- gsub("M","",data_gross)
> data_gross
> data_gross<- substring(data_gross,2,6)
> data_gross
#cek jumlah data_gross
> length(data_gross)
```



Ternyata hanya 86 film yang memiliki nilai gross pendapatan. Sehingga nilai-nilai yang tidak ada tersebut diganti dengan “NA” dengan menggunakan perintah :

```
#Data yang tidak ada diganti dengan nilai "NA"
> for (i in c(49,53,59,60,61,64,64,67,71,72,86,88,92,97)) {
  a<- data_gross[1:(i-1)]
  b<- data_gross[i:length(data_gross)]
  data_gross<- append(a,list("NA"))
  data_gross<- append(data_gross,b)
}
> data_gross
> data_gross<- as.numeric(data_gross)
> data_gross
```

Sehingga akan diperoleh hasil running sebagai berikut

```
Console Terminal Jobs
~/
> gross_data_website<- html_nodes(website, ".ghost~ .text-muted+ span")
> gross_data_website
{xml_nodeset (86)}
[1] <span name="nv" data-value="54,513,740">$54.51M</span>
[2] <span name="nv" data-value="315,058,289">$315.06M</span>
[3] <span name="nv" data-value="176,040,665">$176.04M</span>
[4] <span name="nv" data-value="92,054,159">$92.05M</span>
[5] <span name="nv" data-value="229,024,295">$229.02M</span>
[6] <span name="nv" data-value="63,859,435">$63.86M</span>
[7] <span name="nv" data-value="2,419,031">$2.42M</span>
[8] <span name="nv" data-value="404,515,480">$404.52M</span>
[9] <span name="nv" data-value="168,052,812">$168.05M</span>
[10] <span name="nv" data-value="174,340,174">$174.34M</span>
[11] <span name="nv" data-value="334,201,140">$334.20M</span>
[12] <span name="nv" data-value="389,813,101">$389.81M</span>
[13] <span name="nv" data-value="18,095,701">$18.10M</span>
[14] <span name="nv" data-value="107,825,862">$107.83M</span>
[15] <span name="nv" data-value="327,481,748">$327.48M</span>
[16] <span name="nv" data-value="226,008,385">$226.01M</span>
[17] <span name="nv" data-value="412,563,408">$412.56M</span>
[18] <span name="nv" data-value="17,445,186">$17.45M</span>
[19] <span name="nv" data-value="74,262,031">$74.26M</span>
[20] <span name="nv" data-value="39,175,066">$39.18M</span>
...
> data_gross<- html_text(gross_data_website)
> data_gross
[1] "$54.51M" "$315.06M" "$176.04M" "$92.05M" "$229.02M" "$63.86M" "$2.42M" "$404.52M"
[9] "$168.05M" "$174.34M" "$334.20M" "$389.81M" "$18.10M" "$107.83M" "$327.48M" "$226.01M"
[17] "$412.56M" "$17.45M" "$74.26M" "$39.18M" "$36.25M" "$30.01M" "$188.37M" "$40.44M"
[25] "$209.73M" "$58.06M" "$132.42M" "$504.01M" "$41.19M" "$226.28M" "$33.70M" "$100.23M"
[33] "$172.56M" "$75.47M" "$24.80M" "$85.36M" "$146.88M" "$48.96M" "$102.83M" "$8.04M"
[41] "$33.80M" "$92.03M" "$2.34M" "$620.18M" "$21.02M" "$28.78M" "$17.80M" "$27.78M"
[49] "$130.17M" "$2.03M" "$56.47M" "$81.90M" "$2.29M" "$30.23M" "$18.60M" "$51.69M"
[57] "$20.50M" "$80.10M" "$40.56M" "$18.34M" "$102.09M" "$5.90M" "$21.12M" "$3.48M"
[65] "$2.52M" "$44.31M" "$38.05M" "$29.82M" "$3.02M" "$34.39M" "$50.70M" "$55.68M"
[73] "$104.90M" "$0.10M" "$6.67M" "$10.58M" "$108.25M" "$152.90M" "$51.34M" "$42.87M"
[81] "$1.50M" "$25.11M" "$30.35M" "$0.11M" "$16.79M" "$175.75M"

> data_gross<- gsub("M","",data_gross)
> data_gross
[1] "$54.51" "$315.06" "$176.04" "$92.05" "$229.02" "$63.86" "$2.42" "$404.52" "$168.05"
[10] "$174.34" "$334.20" "$389.81" "$18.10" "$107.83" "$327.48" "$226.01" "$412.56" "$17.45"
[19] "$74.26" "$39.18" "$36.25" "$30.01" "$188.37" "$40.44" "$209.73" "$58.06" "$132.42"
[28] "$504.01" "$41.19" "$226.28" "$33.70" "$100.23" "$172.56" "$75.47" "$24.80" "$85.36"
[37] "$146.88" "$48.96" "$102.83" "$8.04" "$33.80" "$92.03" "$2.34" "$620.18" "$21.02"
[46] "$28.78" "$17.80" "$27.78" "$130.17" "$2.03" "$56.47" "$81.90" "$2.29" "$30.23"
[55] "$18.60" "$51.69" "$20.50" "$80.10" "$40.56" "$18.34" "$102.09" "$5.90" "$21.12"
[64] "$3.48" "$2.52" "$44.31" "$38.05" "$29.82" "$3.02" "$34.39" "$50.70" "$55.68"
[73] "$104.90" "$0.10" "$6.67" "$10.58" "$108.25" "$152.90" "$51.34" "$42.87" "$1.50"
[82] "$25.11" "$30.35" "$0.11" "$16.79" "$175.75"

> data_gross<- substring(data_gross,2,6)
> data_gross
[1] "$4.51" "$15.06" "$16.04" "$2.05" "$29.02" "$3.86" "$2.42" "$40.52" "$16.05"
[13] "$18.10" "$107.8" "$226.0" "$41.5" "$17.45" "$74.26" "$39.18" "$36.25" "$30.01" "$188.3" "$40.44"
[25] "$209.7" "$58.06" "$132.4" "$504.0" "$41.19" "$226.2" "$33.70" "$100.2" "$172.5" "$75.47" "$24.80" "$85.36"
[37] "$146.8" "$48.96" "$102.8" "$8.04" "$33.80" "$92.03" "$2.34" "$620.1" "$21.02" "$28.78" "$17.80" "$27.78"
[49] "$130.1" "$2.03" "$56.47" "$81.90" "$2.29" "$30.23" "$18.60" "$51.69" "$20.50" "$80.10" "$40.56" "$18.34"
[61] "$20.50" "$5.90" "$21.12" "$3.48" "$2.52" "$44.31" "$38.05" "$29.82" "$3.02" "$34.39" "$50.70" "$55.68"
[73] "$104.9" "$0.10" "$6.67" "$10.58" "$108.2" "$152.9" "$51.34" "$42.87" "$1.50" "$25.11" "$30.35" "$0.11"
[85] "$16.79" "$175.7"

>

> #cek jumlah data_gross
> length(data_gross)
[1] 86
> #data yang tidak ada diganti dengan nilai "NA"
> for (i in c(49,53,59,60,61,64,64,67,71,72,86,88,92,97)){
+   a<- data_gross[i:(i-1)]
+   b<- data_gross[i:length(data_gross)]
+   data_gross<- append(a,list("NA"))
+   data_gross<- append(data_gross,b)
+ }
> data_gross
[[1]]
[1] "54.51"

[[2]]
[1] "315.0"

[[3]]
[1] "176.0"

[[4]]
[1] "92.05"

[[5]]
[1] "229.0"

[[6]]
[1] "63.86"

[[7]]
[1] "2.42"

[[8]]
[1] "404.5"

[[9]]
[1] "168.0"
```



```

Console Terminal Jobs
~/
[[92]]
[1] "NA"

[[93]]
[1] "42.87"

[[94]]
[1] "1.50"

[[95]]
[1] "25.11"

[[96]]
[1] "30.35"

[[97]]
[1] "NA"

[[98]]
[1] "0.11"

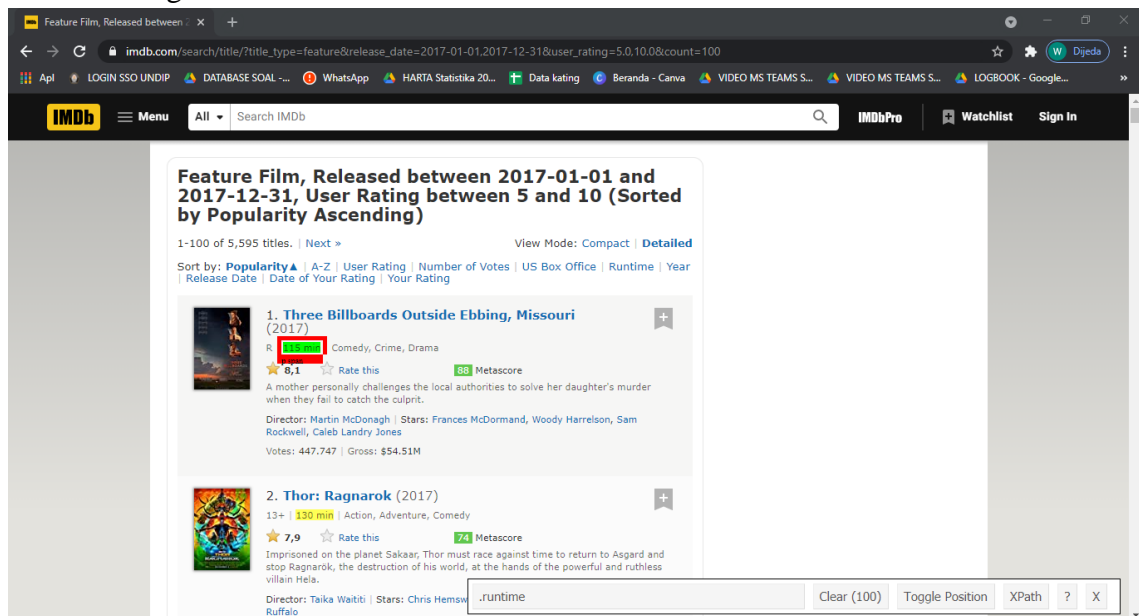
[[99]]
[1] "16.79"

[[100]]
[1] "175.7"

> data_gross<- as.numeric(data_gross)
There were 14 warnings (use warnings() to see them)
> data_gross
 [1] 54.51 315.00 176.00 92.05 229.00 63.86 2.42 404.50 168.00 174.30 334.20 389.80 18.10 107.80 327.40 226.00
[17] 412.50 17.45 74.26 39.18 36.25 30.01 188.30 40.44 209.70 58.06 132.40 504.00 41.19 226.20 33.70 100.20
[33] 172.50 75.47 24.80 85.36 146.80 48.96 102.80 8.04 33.80 92.03 2.34 620.10 21.02 28.78 17.80 27.78
[49] NA 130.10 2.03 56.47 NA 81.90 2.29 30.23 18.60 51.69 NA NA 20.50 80.10 NA
[65] NA 40.56 NA 18.34 102.00 5.90 NA 21.12 3.48 2.52 44.31 38.05 29.82 3.02 34.39
[81] 50.70 55.68 104.90 0.10 6.67 NA 10.58 NA 108.20 152.90 51.34 NA 42.87 1.50 25.11 30.35
[97] NA 0.11 16.79 175.70
>

```

- Yang terakhir mengambil data runtime dari halaman website dengan bantuan SelectorGadget



lalu masukkan perintah ke dalam R seperti berikut :

```

> runtime_data_website <-html_nodes(website, ".runtime")
> runtime_data_website
> data_runtime <-html_text(runtime_data_website)
> data_runtime
> data_runtime <-gsub(" min","",data_runtime)
> data_runtime
> data_runtime <-as.numeric(data_runtime)
> data_runtime

```

Sehingga diperoleh hasil running sebagai berikut


```

Console Terminal Jobs
~/
> runtime_data_website <-html_nodes(website, ".runtime")
> runtime_data_website
{xml_nodeset (100)}
[1] <span class="runtime">115 min</span>
[2] <span class="runtime">130 min</span>
[3] <span class="runtime">104 min</span>
[4] <span class="runtime">164 min</span>
[5] <span class="runtime">120 min</span>
[6] <span class="runtime">123 min</span>
[7] <span class="runtime">104 min</span>
[8] <span class="runtime">119 min</span>
[9] <span class="runtime">118 min</span>
[10] <span class="runtime">105 min</span>
[11] <span class="runtime">133 min</span>
[12] <span class="runtime">136 min</span>
[13] <span class="runtime">132 min</span>
[14] <span class="runtime">113 min</span>
[15] <span class="runtime">135 min</span>
[16] <span class="runtime">136 min</span>
[17] <span class="runtime">141 min</span>
[18] <span class="runtime">127 min</span>
[19] <span class="runtime">122 min</span>
[20] <span class="runtime">126 min</span>
...
> data_runtime <-html_text(runtime_data_website)
> data_runtime
[1] "115 min" "130 min" "104 min" "164 min" "120 min" "123 min" "104 min" "119 min" "118 min"
[10] "105 min" "133 min" "136 min" "132 min" "113 min" "135 min" "136 min" "141 min" "127 min"
[19] "122 min" "126 min" "111 min" "119 min" "106 min" "103 min" "105 min" "116 min" "113 min"
[28] "129 min" "136 min" "137 min" "109 min" "141 min" "129 min" "118 min" "101 min" "124 min"
[37] "140 min" "94 min" "114 min" "107 min" "141 min" "129 min" "118 min" "101 min" "124 min"
[46] "140 min" "121 min" "118 min" "94 min" "154 min" "102 min" "125 min" "116 min" "111 min"
[55] "121 min" "104 min" "100 min" "115 min" "120 min" "103 min" "85 min" "110 min" "111 min"
[64] "123 min" "120 min" "107 min" "93 min" "134 min" "109 min" "111 min" "108 min" "117 min"
[73] "104 min" "114 min" "89 min" "89 min" "92 min" "134 min" "98 min" "113 min" "95 min"
[82] "96 min" "93 min" "108 min" "119 min" "115 min" "93 min" "117 min" "126 min" "102 min"
[91] "115 min" "134 min" "120 min" "151 min" "132 min" "112 min" "105 min" "118 min" "143 min"
[100] "104 min"
...
> data_runtime <-gsub(" min","",data_runtime)
> data_runtime
[1] "115" "130" "104" "164" "120" "123" "104" "119" "118" "105" "133" "136" "132" "113" "135" "136"
[17] "141" "127" "122" "126" "111" "119" "106" "103" "105" "116" "113" "129" "136" "137" "109" "141"
[33] "129" "118" "101" "124" "140" "94" "114" "107" "107" "122" "133" "152" "130" "140" "121" "118"
[49] "94" "154" "102" "125" "121" "116" "121" "104" "100" "115" "120" "103" "85" "110" "111" "123"
[65] "120" "107" "93" "134" "109" "111" "108" "117" "104" "114" "89" "89" "92" "134" "98" "113"
[81] "95" "96" "93" "108" "119" "115" "93" "117" "126" "102" "115" "134" "120" "151" "132" "112"
[97] "105" "118" "143" "104"
...
> data_runtime <-as.numeric(data_runtime)
> data_runtime
[1] 115 130 104 164 120 123 104 119 118 105 133 136 132 113 135 136 141 127 122 126 111 119 106 103
[25] 105 116 113 129 136 137 109 141 129 118 101 124 140 94 114 107 107 122 133 152 130 140 121 118
[49] 94 154 102 125 121 116 121 104 100 115 120 103 85 110 111 123 120 107 93 134 109 111 108 117
[73] 104 114 89 89 92 134 98 113 95 96 93 108 119 115 93 117 126 102 115 134 120 151 132 112
[97] 105 118 143 104
>

```

➤ Statistik Deskriptif

1. Gabungkan semua data menjadi data frame dengan perintah R seperti berikut :

```

> data_film <-data.frame(Runtime = data_runtime,
                        Genre = data_genre, Rating =
                        data_rating,
                        Gross_Pendapatan = data_gross)
> str(data_film)

```

Kemudian akan diperoleh hasil running sebagai berikut

```

Console Terminal Jobs
~/
> data_film <-data.frame(Runtime = data_runtime,
+                        Genre = data_genre, Rating = data_rating,
+                        Gross_Pendapatan = data_gross)
> str(data_film)
'data.frame': 100 obs. of 4 variables:
 $ Runtime      : num  115 130 104 164 120 123 104 119 118 105 ...
 $ Genre        : Factor w/ 9 levels "Action","Adventure"...: 5 1 9 1 1 2 7 1 1 4 ...
 $ Rating       : num  8.1 7.9 7.7 8 6.2 7.3 7.4 6.9 6.6 7.6 ...
 $ Gross_Pendapatan: num  54.5 315 176 92 229 ...
>

```

2. Untuk mencari statistik deskriptifnya dapat menggunakan perintah seperti berikut :

```

> summary(data_film)
#Untuk data runtime
> var(data_runtime)
> sd(data_runtime)
> modus_runtime <-function(data_runtime){
  unik_runtime <-unique(data_runtime)
  tabel_runtime <-table(data_runtime)

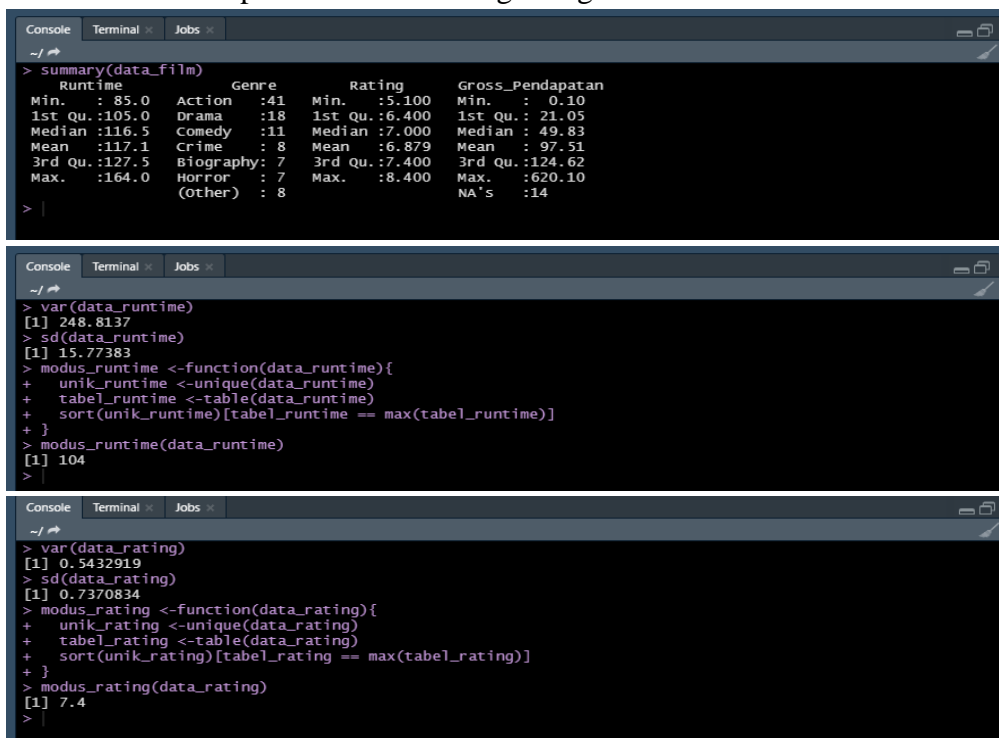
```

```

    sort(unik_runtime)[tabel_runtime == max(tabel_runtime)]
  }
> modus_runtime(data_runtime)
#Untuk data rating
> var(data_rating)
> sd(data_rating)
> modus_rating <-function(data_rating){
  unik_rating <-unique(data_rating)
  tabel_rating <-table(data_rating)
  sort(unik_rating)[tabel_rating == max(tabel_rating)]
}
> modus_rating(data_rating)
#Untuk data gross pendapatan
> var(data_gross)
> sd(data_gross)
> modus_gross <-function(data_gross){
  unik_gross <-unique(data_gross)
  tabel_gross <-table(data_gross)
  sort(unik_gross)[tabel_gross == max(tabel_gross)]
}
> modus_gross(data_gross)

```

Kemudian akan diperoleh hasil running sebagai berikut



The image shows three sequential screenshots of an R terminal window. The first screenshot displays the output of the `summary(data_film)` function, which provides a detailed summary of the 'data_film' dataset, including statistics for Runtime, Genre, Rating, and Gross_Pendapatan. The second screenshot shows the execution of `var(data_runtime)`, `sd(data_runtime)`, and a custom function `modus_runtime` applied to `data_runtime`, resulting in the values 248.8137, 15.77383, and 104 respectively. The third screenshot shows the execution of `var(data_rating)`, `sd(data_rating)`, and the custom function `modus_rating` applied to `data_rating`, resulting in the values 0.5432919, 0.7370834, and 7.4 respectively.

```

> summary(data_film)
  Runtime      Genre      Rating      Gross_Pendapatan
Min.   : 85.0   Action   :41   Min.   :5.100   Min.   : 0.10
1st Qu.:105.0   Drama    :18   1st Qu.:6.400   1st Qu.: 21.05
Median :116.5   Comedy   :11   Median :7.000   Median : 49.83
Mean   :117.1   Crime    : 8   Mean   :6.879   Mean   : 97.51
3rd Qu.:127.5   Biography: 7   3rd Qu.:7.400   3rd Qu.:124.62
Max.   :164.0   Horror   : 7   Max.   :8.400   Max.   :620.10
      (other) : 8      NA's   :14
>
> var(data_runtime)
[1] 248.8137
> sd(data_runtime)
[1] 15.77383
> modus_runtime <-function(data_runtime){
+   unik_runtime <-unique(data_runtime)
+   tabel_runtime <-table(data_runtime)
+   sort(unik_runtime)[tabel_runtime == max(tabel_runtime)]
+ }
> modus_runtime(data_runtime)
[1] 104
>
> var(data_rating)
[1] 0.5432919
> sd(data_rating)
[1] 0.7370834
> modus_rating <-function(data_rating){
+   unik_rating <-unique(data_rating)
+   tabel_rating <-table(data_rating)
+   sort(unik_rating)[tabel_rating == max(tabel_rating)]
+ }
> modus_rating(data_rating)
[1] 7.4
>

```

```

Console Terminal Jobs
~/
> var(data_gross)
[1] NA
> sd(data_gross)
[1] NA
> modus_gross <-function(data_gross){
+   unik_gross <-unique(data_gross)
+   tabel_gross <-table(data_gross)
+   sort(unik_gross)[tabel_gross == max(tabel_gross)]
+ }
> modus_gross(data_gross)
[1] 0.10 0.11 1.50 2.03 2.29 2.34 2.42 2.52 3.02 3.48 5.90 6.67 8.04 10.58
[15] 16.79 17.45 17.80 18.10 18.34 18.60 20.50 21.02 21.12 24.80 25.11 27.78 28.78 29.82
[29] 30.01 30.23 30.35 33.70 33.80 34.39 36.25 38.05 39.18 40.44 40.56 41.19 42.87 44.31
[43] 48.96 50.70 51.34 51.69 54.51 55.68 56.47 58.06 63.86 74.26 75.47 80.10 81.90 85.36
[57] 92.03 92.05 100.20 102.00 102.80 104.90 107.80 108.20 130.10 132.40 146.80 152.90 168.00 172.50
[71] 174.30 175.70 176.00 188.30 209.70 226.00 226.20 229.00 315.00 327.40 334.20 389.80 404.50 412.50
[85] 504.00 620.10
>

```

➤ Output Hasil

- Data Frame Film

Karena sebelumnya data film sudah dibuat data frame maka perintah untuk melihat data frame film menggunakan `> data_film`

```

Console Terminal Jobs
~/
> #Tabel data_film
> data_film
  Runtime Genre Rating Gross_Pendapatan
1      115 Comedy   8.1      54.51
2      130 Action   7.9     315.00
3      104 Horror   7.7     176.00
4      164 Action   8.0     92.05
5      120 Action   6.2    229.00
6      123 Adventure 7.3     63.86
7      104 Drama    7.4       2.42
8      119 Action   6.9    404.50
9      118 Action   6.6    168.00
10     105 Biography 7.6    174.30
11     133 Action   7.4    334.20
12     136 Action   7.6    389.80
13     132 Drama    7.9    18.10
14     113 Action   7.6    107.80
15     135 Horror   7.3    327.40
16     136 Action   6.6    226.00
17     141 Action   7.4    412.50
18     127 Biography 7.0    17.45
19     122 Horror   6.4     74.26
20     126 Action   6.7     39.18
21     111 Action   6.2     36.25
22     119 Biography 7.5     30.01
23     106 Action   7.8    188.30
24     103 Animation 7.8     40.44
25     105 Animation 8.4    209.70
26     116 Action   5.5     58.06
27     113 Drama    8.0    132.40
28     129 Family   7.1    504.00
29     136 Action   6.5    41.19
30     137 Action   8.1    226.20
31     109 Action   5.3     33.70
32     141 Action   6.7    100.20
33     129 Action   6.5    172.50
34     118 Action   6.9     75.47
35     101 Drama    7.6     24.80
36     124 Action   5.9     85.36
37     140 Action   7.4    146.80
38     94 Comedy   7.4     48.96
39     114 Crime    6.5    102.80
40     107 Comedy   7.2     8.04
41     107 Crime    7.7    33.80
42     122 Action   7.5    92.03
43     133 Adventure 7.2     2.34
44     152 Action   7.0    620.10
45     130 Drama    7.5    21.02
46     140 Biography 7.4    28.78
47     121 Drama    6.6    17.80
48     118 Comedy   7.0    27.78
49     94 Horror    6.3       NA
50     154 Action   5.2    130.10
51     102 Crime    7.4     2.03
52     125 Biography 7.4    56.47
53     121 Crime    7.3       NA
54     116 Drama    7.2    81.90
55     121 Drama    7.0     2.29
56     104 Horror    6.6    30.23
57     100 Action   6.0    18.60
58     115 Action   6.7    51.69
59     120 Action   6.4       NA
60     103 Drama    6.5       NA
61     85 Comedy    6.3       NA
62     110 Drama    5.4    20.50
63     111 Action   5.4    80.10
64     123 Action   6.9       NA
65     120 Action   7.3       NA
66     107 Action   6.3    40.56
67     93 Action   5.8       NA
68     134 Action   7.6    18.34
69     109 Horror    6.5    102.00
70     111 Drama    7.6     5.90
71     108 Horror    7.4       NA
72     117 Action   6.3       NA
73     104 Biography 7.4    21.12
74     114 Drama    6.6     3.48
75     89 Crime     6.8     2.52
76     89 Adventure 5.7    44.31
77     92 Crime     5.8    38.05
78     134 Drama    7.2    29.82
79     98 Comedy    6.6     3.02
80     113 Action   7.0    34.39
81     95 Action   5.6    50.70
82     96 Comedy    6.6    55.68
83     93 Comedy    5.8    104.90
84     108 Action   6.4     0.10
85     119 Crime    5.1     6.67
86     115 Action   6.7       NA
87     93 Drama     6.3    10.58
88     117 Drama    6.3       NA
89     126 Comedy   7.0    108.20
90     102 Animation 6.7    152.90
91     115 Action   7.2    51.34
92     134 Drama    7.4       NA
93     120 Comedy   7.5    42.87
94     151 Comedy   7.2     1.50
95     132 Biography 6.8    25.11
96     112 Action   6.4    30.35
97     105 Drama    5.2       NA
98     118 Drama    7.4     0.11
99     143 Crime     7.3    16.79
100    104 Animation 7.3    175.70
>

```

- Histogram Data Film

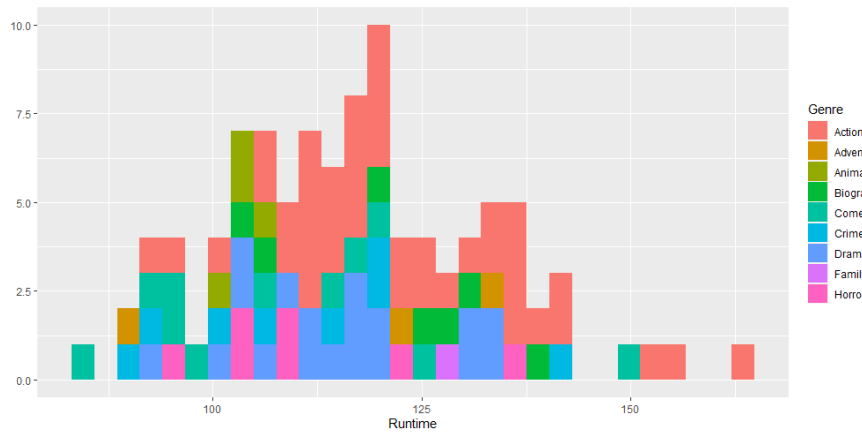
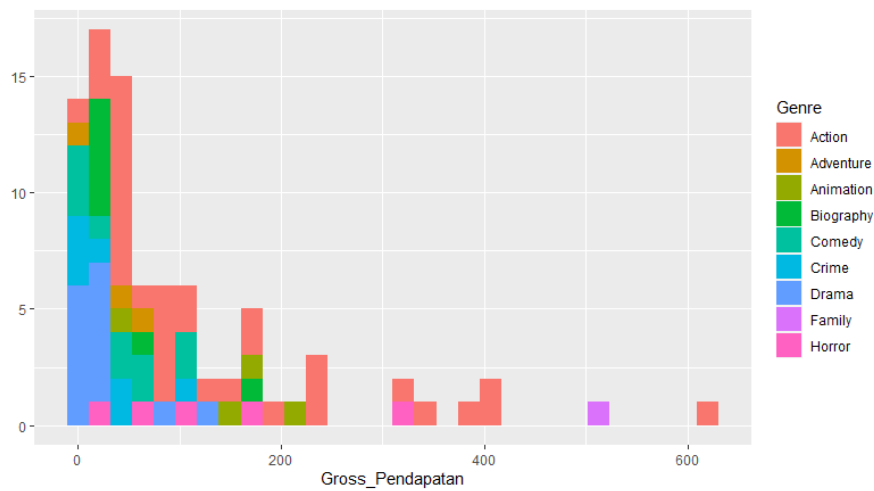
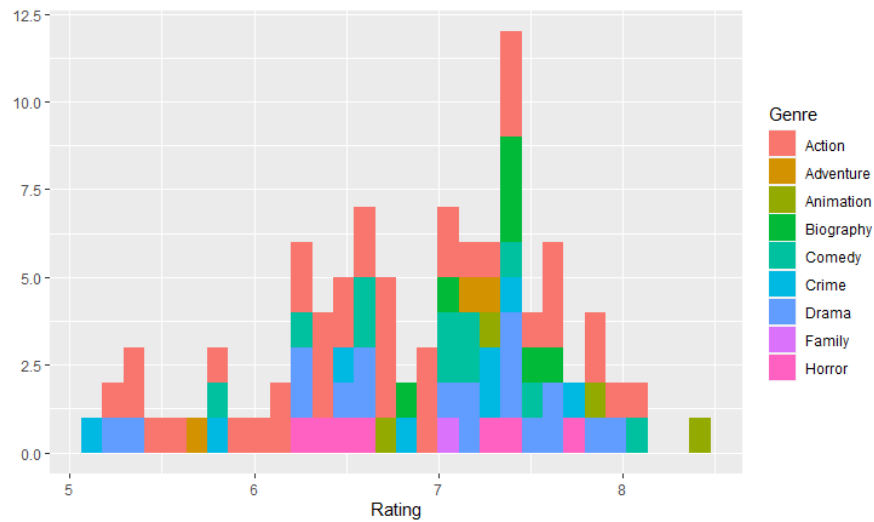
Perintah dalam R untuk membuat histogram data film yaitu

```

> library("ggplot2")
> qplot(data = data_film, Rating, fill = Genre, bins = 30)
> qplot(data = data_film, Gross_Pendapatan, fill =
Genre, bins = 30)
> qplot(data = data_film, Runtime, fill = Genre, bins = 30)

```

Kemudian akan diperoleh hasil output sebagai berikut

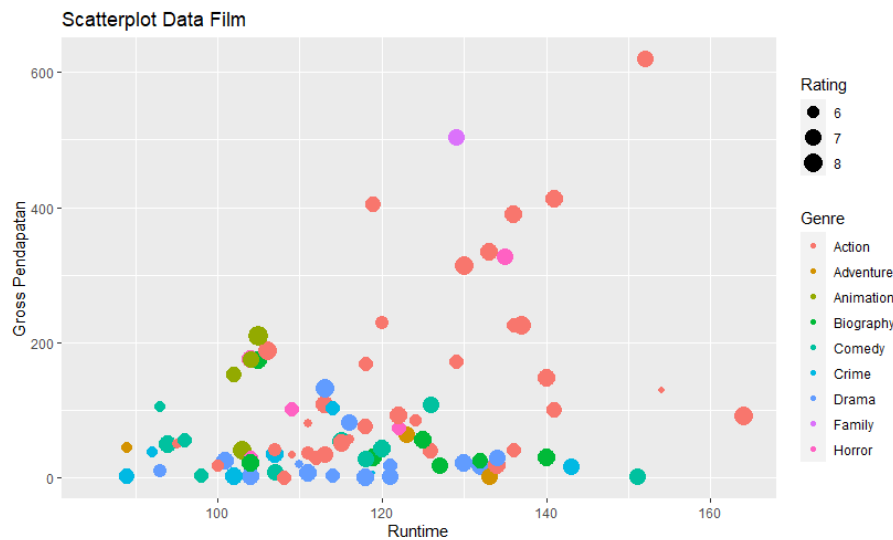


- **Scatterplot Data Film**

Perintah dalam R untuk membuat scatterplot data film yaitu

```
> ggplot(data_film, aes(x=Runtime, y=Gross_Pendapatan)) +  
  geom_point(aes(size=Rating, col=Genre)) +  
  labs(title = "Scatterplot Data Film",  
        y = "Gross Pendapatan")
```

Kemudian akan diperoleh hasil output sebagai berikut



➤ Analisis Data

- Statistik Deskriptif

Berdasarkan hasil analisis statistik deskriptif menggunakan R diperoleh :

1. Nilai terkecil atau minimum dari data Runtime, Rating, dan Gross_Pendapatan masing-masing adalah 85.0, 5.100, dan 0.10.
2. Nilai terbesar atau maksimum dari data Runtime, Rating, dan Gross_Pendapatan masing-masing adalah 164.0, 8.400, dan 620.10.
3. Kuartil bawah (1st Qu.) dari data Runtime, Rating, dan Gross_Pendapatan masing-masing adalah 105.0, 6.400, dan 21.05.
4. Median dari data Runtime, Rating, dan Gross_Pendapatan masing-masing adalah 116.5, 7.000, dan 49.83.
5. Rata-rata atau mean dari data Runtime, Rating, dan Gross_Pendapatan masing-masing adalah 117.1, 6.879, dan 97.51.
6. Kuartil atas (3st Qu.) dari data Runtime, Rating, dan Gross_Pendapatan masing-masing adalah 127.5, 7.400, dan 124.62.
7. Varian dari data Runtime dan Rating masing-masing adalah 248.8137 dan 0.5432919, sedangkan untuk data Gross_Pendapatan bernilai NA karena terdapat beberapa data yang tidak ada nilainya.
8. Standar deviasi dari data Runtime dan Rating masing-masing adalah 15.77383 dan 0.7370834, sedangkan untuk data Gross_Pendapatan bernilai NA karena terdapat beberapa data yang tidak ada nilainya.
9. Modus dari data Runtime dan Rating masing-masing adalah 104 dan 7.4, sedangkan untuk data Gross_Pendapatan tidak mempunyai modus karena semua data hanya muncul 1 kali.
10. Data Genre tidak mempunyai nilai terkecil, nilai terbesar, kuartil bawah, kuartil atas, median, mean, varian, dan standar deviasi karena data Genre termasuk data kualitatif. Tetapi data Genre mempunyai modus yaitu Action sehingga genre film yang banyak disukai pada tahun 2017 adalah Action.

- Histogram

Berdasarkan hasil output histogram dapat diketahui bahwa 100 film teratas berdasarkan popularitas pada tahun 2017 dengan genre Animation mempunyai rating paling tinggi sedangkan film dengan genre Crime mempunyai rating paling rendah. Film dengan genre Action juga mempunyai gross pendapatan paling tinggi dibandingkan genre lainnya. Dan rata-rata runtime film pada tahun 2017 berada diantara 100 sampai 150 menit tetapi terdapat beberapa film yang mempunyai runtime lebih dari 150 menit yaitu film dengan genre Comedy dan Action.

- Scatterplot

Berdasarkan hasil output scatterplot dapat diketahui bahwa 100 film teratas berdasarkan popularitas pada tahun 2017 mempunyai gross pendapatan film paling tinggi yaitu diatas 600 dengan genre film yaitu Action. Kemudian diurutan kedua adalah film dengan genre Family yaitu sekitar 500 dan rata-rata gross pendapatan film pada tahun 2017 adalah antara 0 sampai 200.