

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





# BÁO CÁO

## ĐỒ ÁN (TIỂU LUẬN, BÀI TẬP LỚN)

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

Phân tích dữ liệu R

***Đề tài:*** Phân tích dữ liệu phim IMDb

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày 8 tháng 1 năm 2022



ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





# BÁO CÁO

## ĐỒ ÁN (TIỂU LUẬN, BÀI TẬP LỚN)

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

Phân tích dữ liệu với R

***Đề tài:*** Phân tích dữ liệu phim IMDb

###### Giảng viên hướng dẫn: Hồ Quốc Dũng

###### Sinh viên thực hiện: Nguyễn Trịnh Tấn Đạt

###### Lớp: KHDL&TTNT

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày 8 tháng 1 năm 2022

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ



PHIẾU ĐÁNH GIÁ ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN

**Học kỳ I, năm học 2021 - 2022**

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1** | **Cán bộ chấm thi 2** |
| **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT1:**  Bằng số: .........................................  Bằng chữ: ....................................... | **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT2:**  Bằng số: .........................................  Bằng chữ: ....................................... |

Điểm kết luận: ...........................................................................................................

Bằng số:.....................................................................................................................

Bằng chữ: .................................................................................................................

*Thừa Thiên Huế, ngày tháng năm 2021*

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1**  *(Ký và ghi rõ họ và tên)* | **Cán bộ chấm thi 2**  *(Ký và ghi rõ họ và tên)* |

**MỤC LỤC**

[CHƯƠNG I: GIỚI THIỆU NGÔN NGỮ R 6](#_Toc92648319)

[I: Khái quát chung 6](#_Toc92648320)

[1. Giới thiệu R 6](#_Toc92648321)

[2. Ưu điểm của R 6](#_Toc92648322)

[II: Hướng dẫn sử dụng R 6](#_Toc92648323)

[1. Cài đặt và giao diện 6](#_Toc92648324)

[2. Nhập dữ liệu trong R 7](#_Toc92648325)

[2.1 Nhập dữ liệu trực tiếp từ dòng lệnh theo cấu trúc từ hàm c(): 7](#_Toc92648326)

[2.2 Nhập dữ liệu từ cửa sổ nhập Data Editor: 8](#_Toc92648327)

[2.3 Nhập dữ liệu từ File text 8](#_Toc92648328)

[2.4 Nhập dữ liệu từ File Excel 8](#_Toc92648329)

[3. Các lệnh hệ thống 8](#_Toc92648330)

[3.1 Lệnh về môi trường vận hành của R 8](#_Toc92648331)

[3.2 Lệnh cơ bản 9](#_Toc92648332)

[3.3 Trợ giúp trong R 9](#_Toc92648333)

[4. Tổ chức dữ liệu trong R 9](#_Toc92648334)

[5. Các hàm thống kê và đồ thị 9](#_Toc92648335)

[5.1 Các hàm thống kê 9](#_Toc92648336)

[5.2 Phân phối thống kê 10](#_Toc92648337)

[5.3 Phân tích thống kê 10](#_Toc92648338)

[5.4 Đồ thị 11](#_Toc92648339)

[CHƯƠNG II: ĐỒ ÁN PHÂN TÍCH DỮ LIỆU VỚI R 12](#_Toc92648340)

[I: Mổ tả tổng quan về dữ liệu được sử dụng 12](#_Toc92648341)

[II: Phân tích thông kê mô tả dữ liệu 12](#_Toc92648342)

[III: Phân tích số liệu 22](#_Toc92648343)

1. GIỚI THIỆU NGÔN NGỮ R
   1. Khái quát chung
      1. Giới thiệu R

Năm 1996, trong một bài báo quan trọng về tính toán thống kê, hai nhà thống kê học Ross Ihaka và Robert Gentlan thuộc Trường đại học Auckland, New Zealand phác hoạ một ngôn ngữ mới cho phân tích thống kê mà họ đặt tên là R. Sáng kiến này được rất nhiều nhà thống kê học trên thế giới tán thành và tham gia vào việc phát triển R.

Vậy R là gì? Nói một cách ngắn gọn, R là một phần mềm sử dụng cho phân tích thống kê và vẽ biểu đồ. Thật ra, về bản chất, R là ngôn ngữ máy tính đa năng, có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí(recreational mathatics) , tính toán ma trận(matrix) , đến các phân tích thống kê phức tạp. Vì là một ngôn ngữ, cho nên người ta có thể sử dụng R để phát triển thành các phần mềm chuyên môn cho một vấn đề tính toán cá biệt.

* + 1. Ưu điểm của R

R có chứa nhiều loại kỹ thuật thống kê: mô hình hóa tuyến tính và phi tuyến, kiểm thử thống kê cổ điển, phân tích chuỗi thời gian, phân loại, phân nhóm, v. v. và đồ họa. R

R cũng có tính mở rộng cao bằng cách sử dụng các gói cho người dùng đưa lên cho một số chức năng và lĩnh vực nghiên cứu cụ thể.

Một điểm mạnh khác của R là nền tảng đồ họa có thể tạo ra những đồ thị chất lượng cao cùng các biểu tượng toán học.

Dù R được dùng chủ yếu bởi những nhà thống kê và cũng có thể dùng làm một công cụ tính toán ma trận tổng quát với các kết quả đo đạc cạnh tranh so với GNU Octave và đối thủ thương mại của nó, MATLAB. Giao diện RWeka đã được thêm vào phần mềm khai phá dữ liệu phổ biến Weka, cho phép đọc/ghi định dạng arff vì vậy cho phép sử dụng tính năng khai phá dữ liệu trong Weka và thống kê trong R

Ngôn ngữ R có rất nhiều ưu điểm so với các ngôn ngữ lập trình bậc cao như C , C++ , Java….

R có khả năng điều khiển dữ liệu và lưu trữ số liệu, R còn có tính nguyên bản.

R cho phép sử dụng ma trận đại số.

Có thể sử dụng bảng băm và các biểu thức chính quy

R cũng hỗ trợ lập trình hướng đối tượng.

Khả năng biểu diễn đồ họa phong phú.

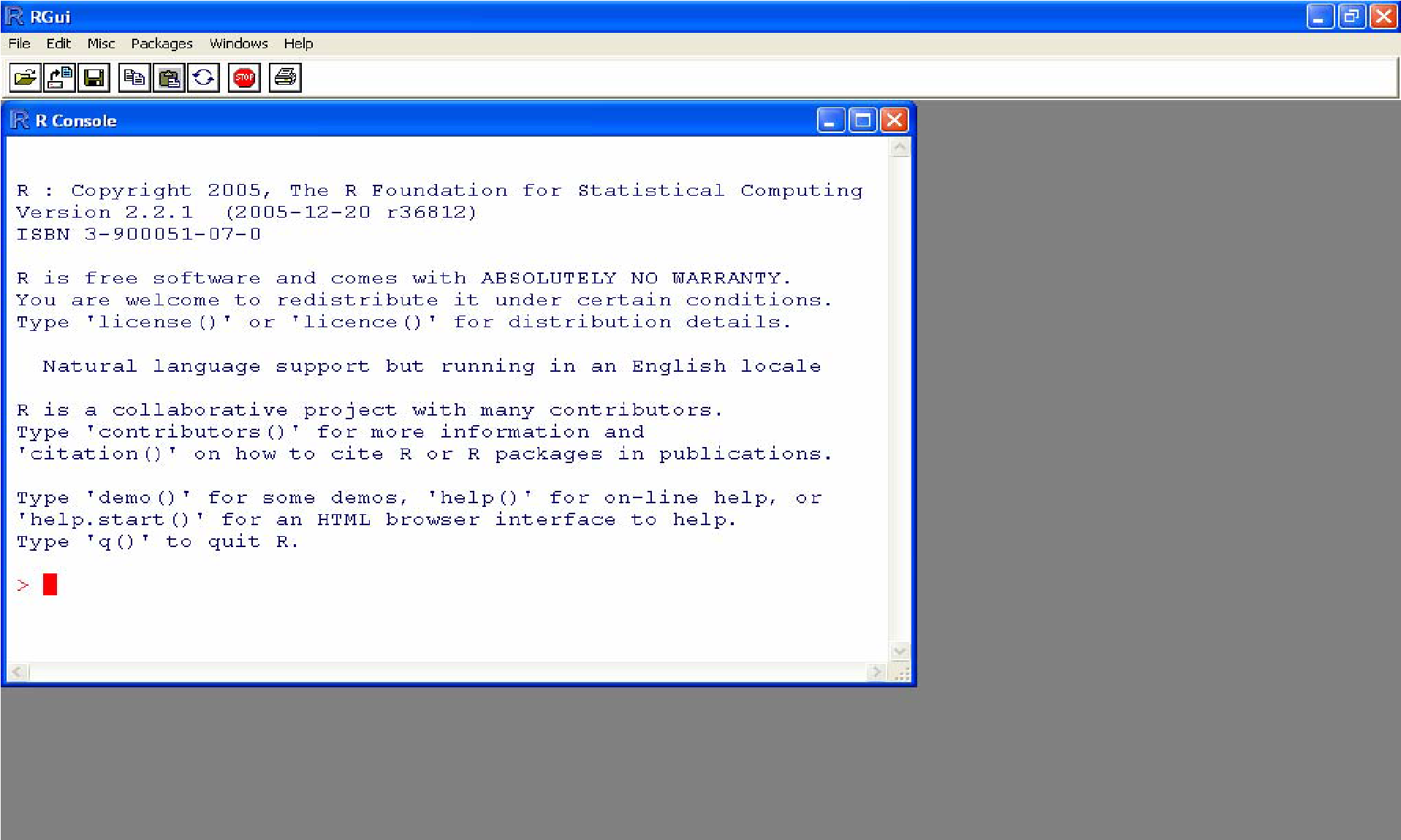
Ngôn ngữ R cũng cung cấp các cấu trúc điều khiển cơ bản như các ngôn ngữ lập trình bậc cao khác. Ví dụ như :If…else…;while…. ;for……vv.

* 1. Hướng dẫn sử dụng R
     1. Cài đặt và giao diện

Để sử dụng R, việc đầu tiên là phải cài đặt R trong máy.

Khi đã tải R xuống máy tính, bước kế tiếp là cài đặt vào máy tính. Để làm việc này, chỉ đơn giản nhấn chuột vào tài liệu trên và làm theo hướng dẫn cách cài đặt trên màn hình. Đây là một bước rất đơn giản, chỉ cần 1 phút là việc cài đặt R có thể hoàn tất.

Sau khi hoàn tất việc cài đặt, một *icon s*ẽ xuất hiện trên *desktop* của máy tính(Hình bên). Đến đây thì đã sẵn sàng sử dụng R. Cóthể nhấp chuột vào icon này và sẽ có một cửa sổnhư sau:



* + 1. Nhập dữ liệu trong R

Dữ liệu mà R hiểu được phải là dữ liệu trong một data.frame.

* + - 1. Nhập dữ liệu trực tiếp từ dòng lệnh theo cấu trúc từ hàm c():

*Tên\_biến\_lưu\_dữ\_liệu <- c(pt1, pt2, …, ptn)*

 VD: a <- c(4,67,87,4,5,3)

b <- c(9,8,7,5,6,5,22)

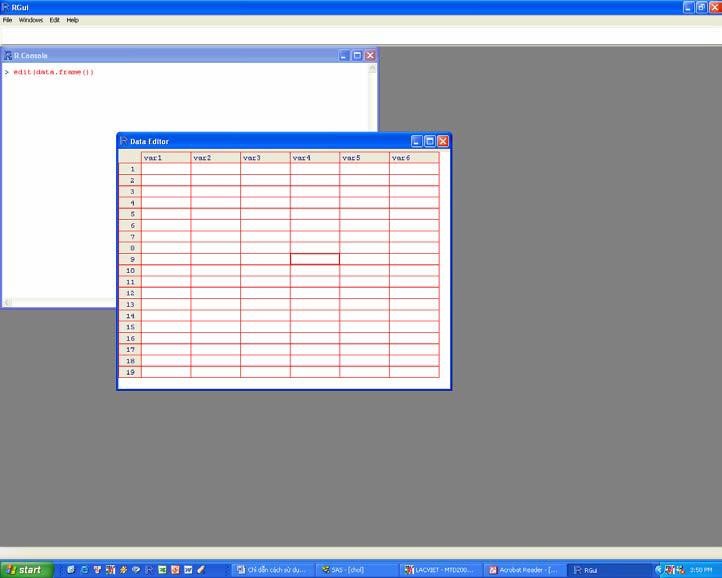
Ghép các biến riêng lẻ nhập bằng hàm c thành một khung dữ liệu để sử dụng sau này:

*Tên\_biến <- data.frame(tham\_số\_1,tham\_số\_2, ……. , tham\_số\_n)*

Lưu tên tệp: *save(tên\_biến ,file=“tênfile.rdata”)*

* + - 1. Nhập dữ liệu từ cửa sổ nhập Data Editor:

**> e***dit(data.frame())*



* + - 1. Nhập dữ liệu từ File text
* Tênbiến <- read.table(“path file", header=TRUE)

Xem lại nội dung tệp vừa nhập:

* names(Tênbiến)

Lưu lại dưới dạng tệp R để xử lý sau này

* *save(tên\_biến ,file=“tênfile.rdata”)* 
  + - 1. Nhập dữ liệu từ File Excel

Lưu tệp excel dưới đuôi \*.csv - Đọc tệp csv với cú pháp sau:

* *Tênbiến <- read.csv(“đường\_dẫn\_đến\_file\_csv”,HEADER=true)*

Tham số *HEADER = true* cho R biết dòng đầu tiên của file xls được chọn làm tên của các cột.

Lưu lại tệp để sử dụng lần sau bằng lệnh save().

* + 1. Các lệnh hệ thống
       1. Lệnh về môi trường vận hành của R

|  |  |
| --- | --- |
| getwd() | Cho biết thư mục hiện hành là gì |
| setwd(c:/works) | Chuyển thư mục vận hành về c:\works(chú ý R dùng |
| options(prompt=”R>”) | Đổi prompt thành R> |
| options(width=100) | Đổi chiều rộng cửa sổ R thành 100 characters |
| options(scipen=3) | Đổi số thành 3 số thập phần(thay vì kiểu 1. 2E-04) |
| options() | Cho biết các thông số về môi trường của R |

* + - 1. Lệnh cơ bản

|  |  |
| --- | --- |
| ls() | Liệt kê các đối tượng trong bộ nhớ |
| rm(object) | Xóa bỏ đối tượng |
| seach() | Tìm hướng |

* + - 1. Trợ giúp trong R

Ngoài lệnh ***args()*** R còn cung cấp lệnh ***help()*** để người sử dụng có thể hiểu “Văn phạm” của từng hàm. Chẳng hạn như muốn biết hàm ***lm*** có những tham số gì chỉ cần gõ lệnh: *>help()*  hay *>?lm*

một cửa sổ sẽ hiện ra bên ngoài của màn hình chỉ rõ cách sử dụng ra sao và thậm

chí có cả ví du.

Sử dụng lệnh *help. start()* một cửa sổ sẽ xuất hiện chỉ dẫn toàn bộ hệ thống R.

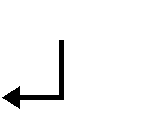
Hàm ***apropos*** cũng rất có ích vì nó cung cấp cho tất cả các hàm trong R bắt đầu bằng kí tự mà muốn tìm. Chẳng hạn như muốn biết hàm nào trong R có kí tự “lm” thì chỉ gõ lệnh:

*> apropos(lm)*

* + 1. Tổ chức dữ liệu trong R

Sử dụng R cho các phép tính ma trận

*Nhập dữ liệu vào ma trận theo cú pháp:*

>tenbien <- Matrix(biendl, nrow)

VD: ta có ma trận A có dạng A =

Khi nhập vào R sẽ nhập như sau:

> x <- c(4,5,6,7,8,9,10,11,12)

> A <- matrix(x, nrow=3) Cho 2 ma trận A và B:

* Cộng (trừ) 2 ma trận: > A+ (-)B
* Nhân 2 ma trận: > A %\*%B
* Ma trận nghịch đảo: > solve(A)
* Ngoài ra R có một gói Matrix chuyên thiết kế cho tính toán ma trận.
  + 1. Các hàm thống kê và đồ thị
       1. Các hàm thống kê

|  |  |
| --- | --- |
| min(x) | Số nhỏ nhất của biến số x |
| max(x) | Số lớn nhất của biến số x |
| which. max(x) | Tìm dòng nào có giá trị lớn nhất của biến số x |
| which. min(x) | Tìm dòng nào có giá trị nhỏ nhất của biến số x |
| sum(x) | Số tổng của biến số x |
| range(x) | Khác biệt giữa max(x) va min(x) |
| mean(x) | Số trung bình của biến số x |
| median(x) | Số trung vị(median) của biến số x |
| sd(x) | Độ lệch chuẩn(standard deviation) của biến số x |
| var(x) | Phương sai(variance) của biến số x |

* + - 1. Phân phối thống kê

|  |  |
| --- | --- |
| pnorm(x, mean, sd) | Phân phối chuẩn |
| plnorm(x, mean, sd) | Phân phối chuẩn logarit |
| pt(x, df) | Phân phối t |
| pf(x, n1, n2) | Phân phối F |

* + - 1. Phân tích thống kê

|  |  |
| --- | --- |
| t. test | Kiểm định t |
| pairwise. t. test | Kiểm định t cho paired design |
| var. test | Kiểm định phương sai |
| bartlett. test | Kiểm định nhiều phương sai |
| wilcoxon. test | Kiểm định Wilcoxon |
| kruskal. test | Kiểm định Kruskal |
| friedman. test | Kiểm định Friedman |
| lm(y ~ x) | Phân tich hồi qui tuyến tính(linear regression) |

* + - 1. Đồ thị

|  |  |
| --- | --- |
| plot(y~x) | Vẽ đồ thị y va x(scatter plot) |
| hist(x) | Vẽ đồ thị y va x(scatter plot) |
| plot(y ~ x | z) | Vẽ hai biểu đồ x va y theo từng nhóm của z |
| pie(x) | Vẽ đồ thị tròn |
| boxplot(x) | Vẽ đồ thị theo dạng hình hộp |
| qqnorm(x) | Vẽ phân phối quantile của biến số x |
| qqplot(x, y) | Vẽ phân phối quantile của biến số y theo x |
| barplot(x) | Vẽ biểu đồ hình khối cho biến số x |
| hist(x) | Vẽ histogram cho biến số x |
| stars(x) | Vẽ biểu đồ sao cho biến số x |
| abline(a, b) | Vẽ đường thẳng với intercept=a va slope=b |
| abline(h=y) | Vẽ đường thẳng ngang |
| abline(v=x) | Vẽ đường thẳng đứng |
| abline(lm. object) | Vẽ đồ thị theo mô hinh tuyến tính |
| pch | Kí hiệu để vẽ đồ thị(pch = *p*lotting *ch*aracters) |
| mfrow, mfcol | Tạo ra nhiều cửa sổ để vẽ nhiều đồ thị |
| xlim, ylim | Cho giới hạn của trục hòanh và trục tung |
| xlab, ylab | Viết tên trục hòanh và trục tung |
| lty, lwd | Dạng và kích thước của đường biểu diễn |
| cex, mex | Kích thước và khoảng cách giữa các kí tự. |
| col | Màu sắc |

1. ĐỒ ÁN PHÂN TÍCH DỮ LIỆU VỚI R
   1. Mổ tả tổng quan về dữ liệu được sử dụng

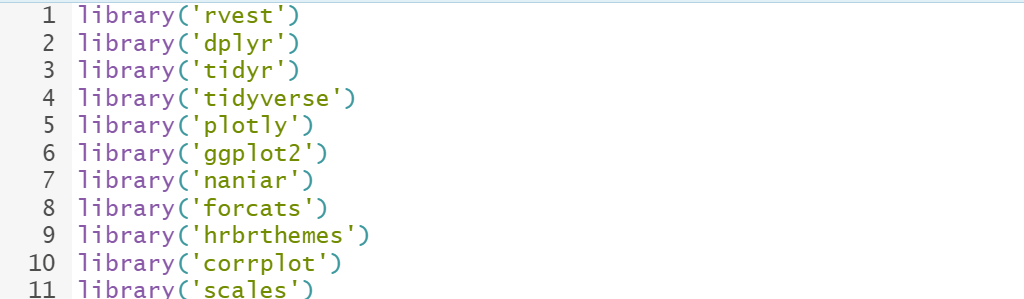
Dữ liệu crawl về liên quan về các phim được sản xuất năm 2019.

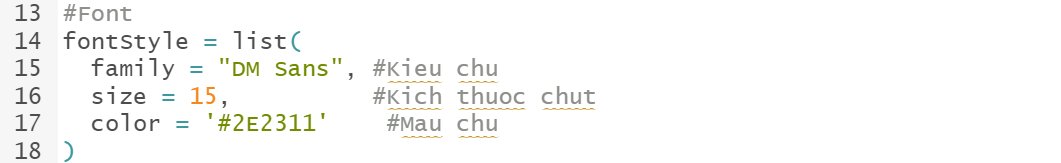
Dữ liệu bao gồm 10 cột, 2000 dòng.

|  |  |
| --- | --- |
| Tên Cột | Kiểu dữ liệu |
| STT | Num |
| Tên phim | Chr |
| Mô tả | Chr |
| Thời gian | Num |
| Thể loại | Factor |
| Sao | Num |
| Điểm số | Chr |
| Bình chọn | Num |
| Đạo diễn | Factor |
| Tổng doanh thu | Factor |

* 1. Phân tích thông kê mô tả dữ liệu

Các thư viện được sử dụng

Thay đổi kiểu chữ trong các đồ thị



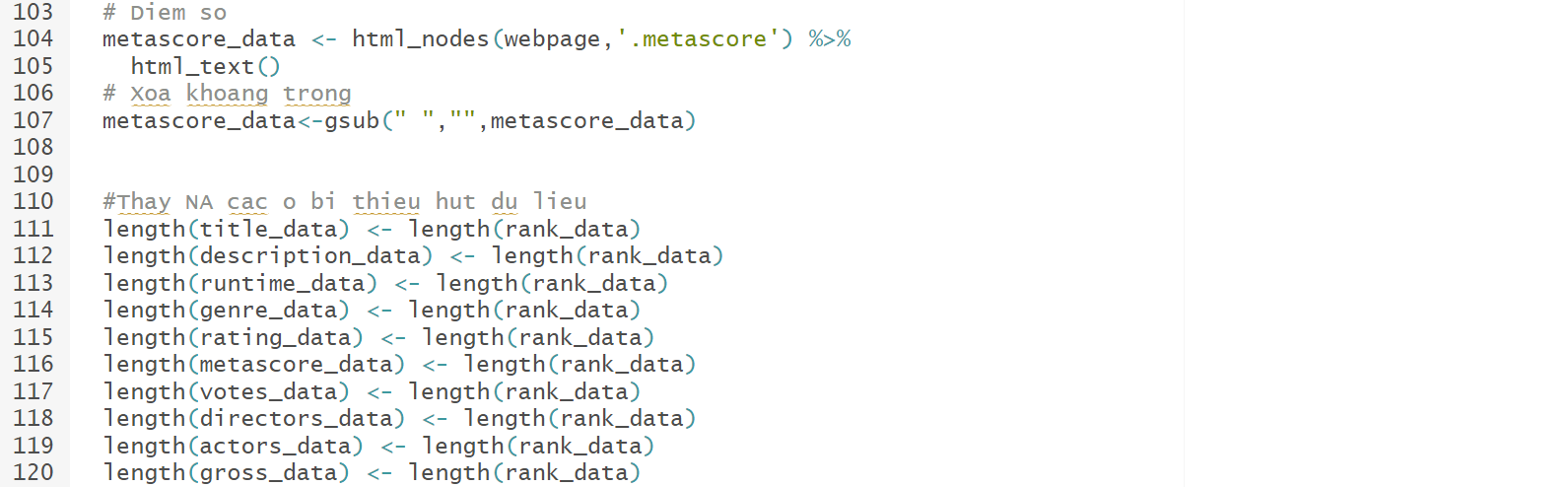
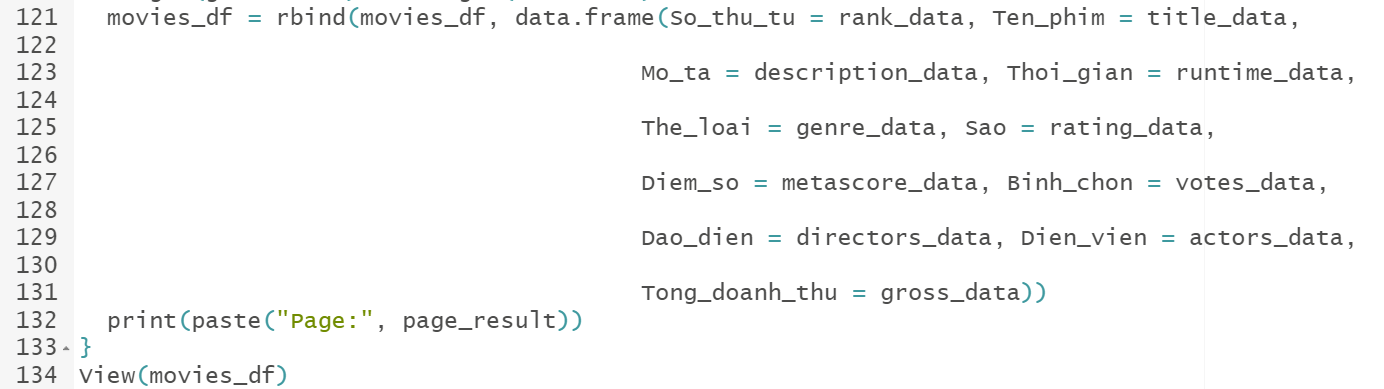
Tạo một dataframe trống

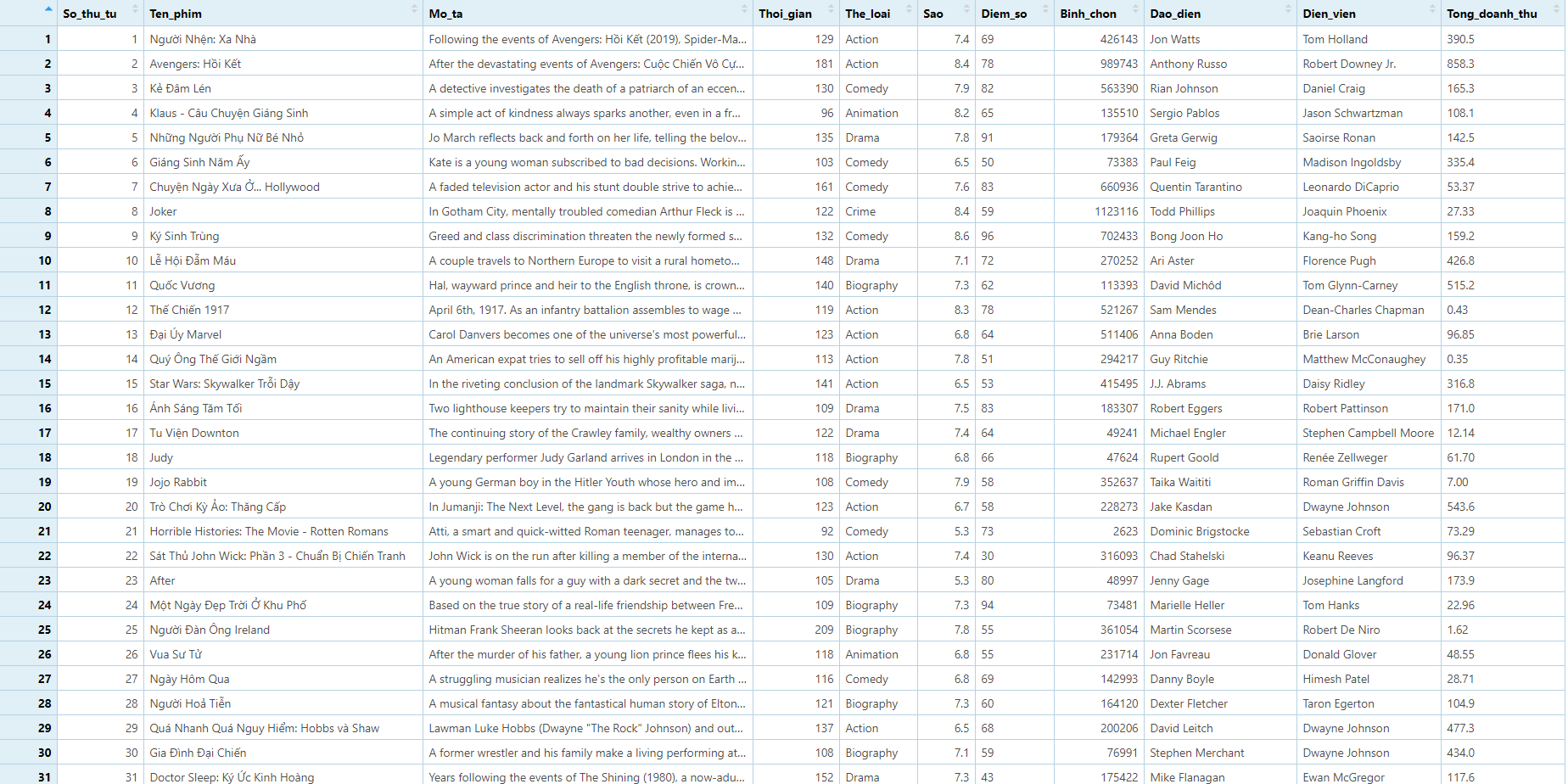


Code tải dữ liệu từ trang web về và bảng dữ liệu thu được



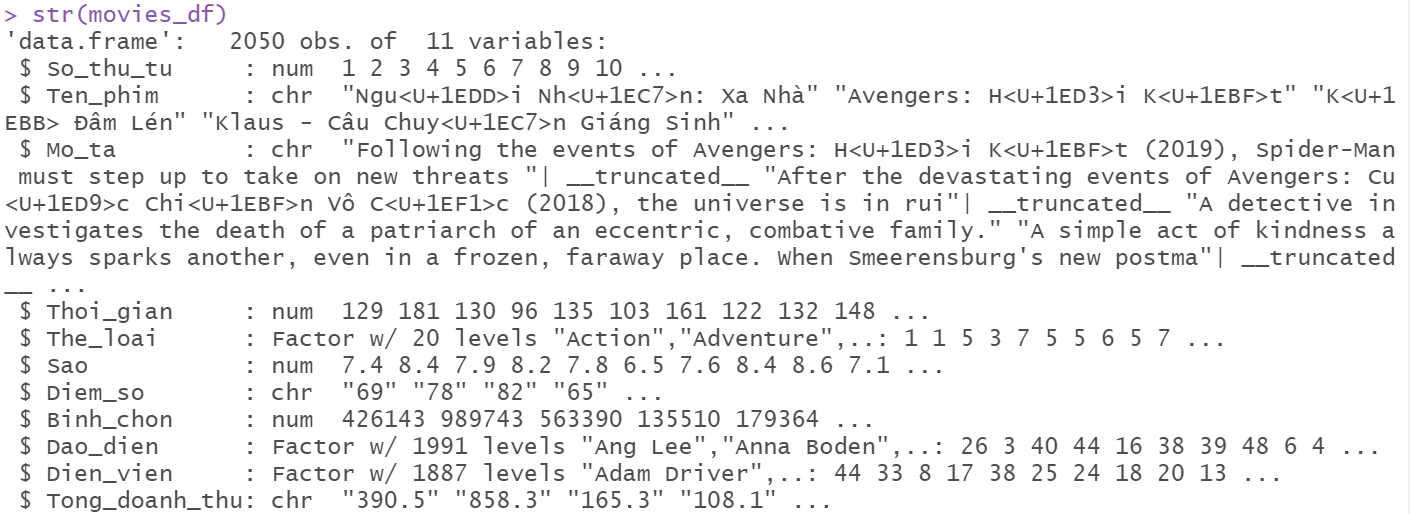


  bảng dữ liệu movies\_df



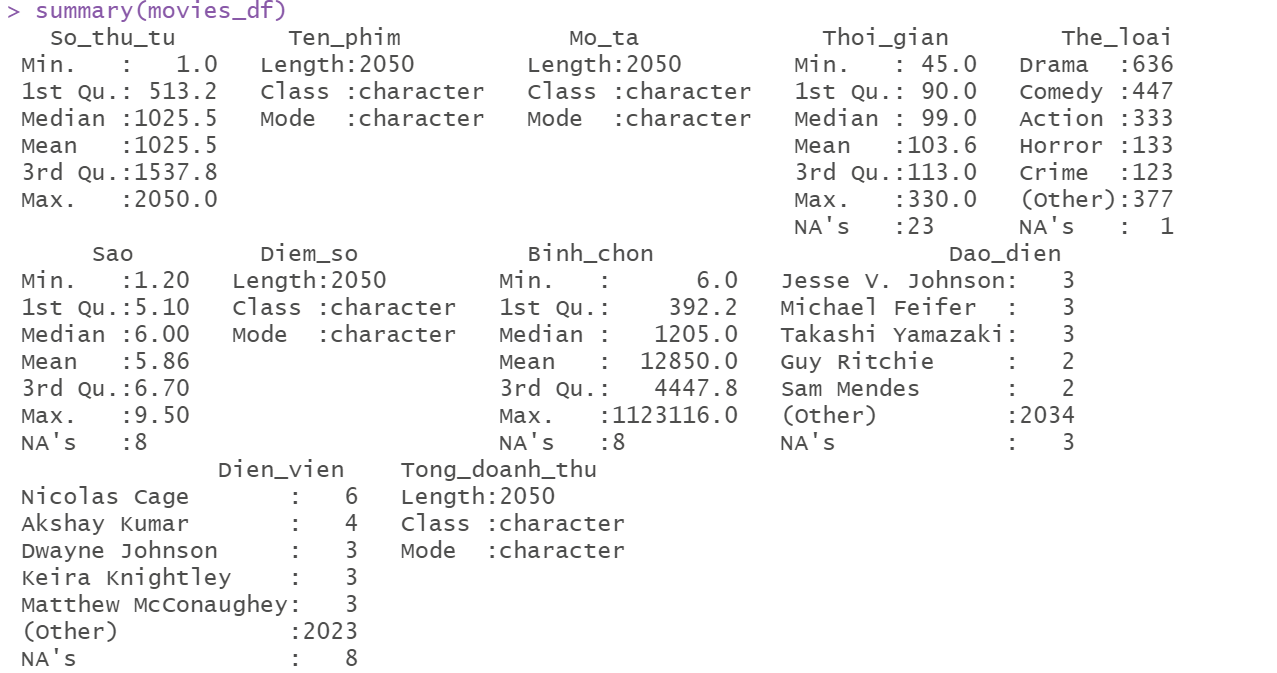
Xem các các kiểu dữ liệu



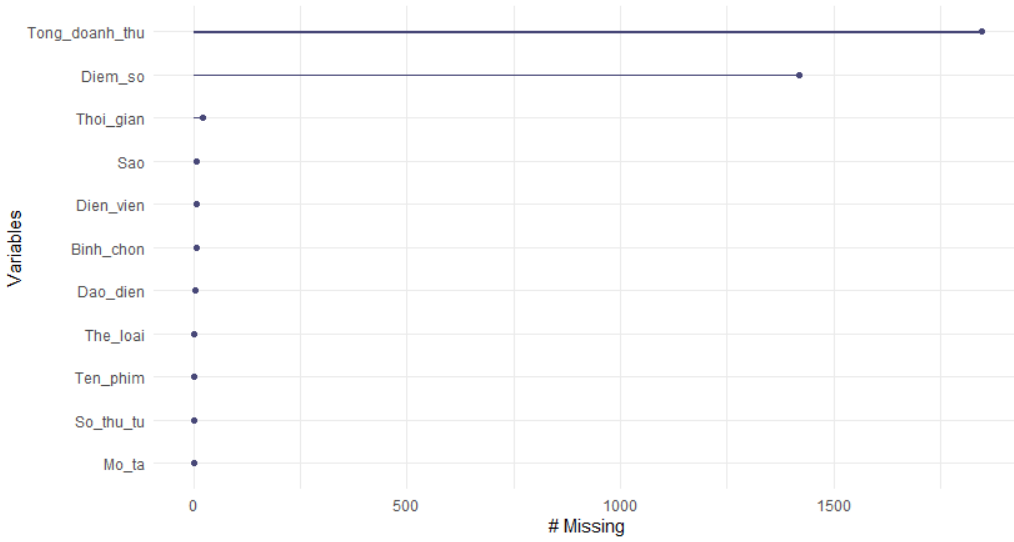


Tính mean, median,.. cho các trường dữ liệu





Biểu đồ hình dung các giá trị thiếu hụt trong dữ liệu



Tạo 1 bảng dữ liệu “imdb” loại bỏ cá giá trị thiếu



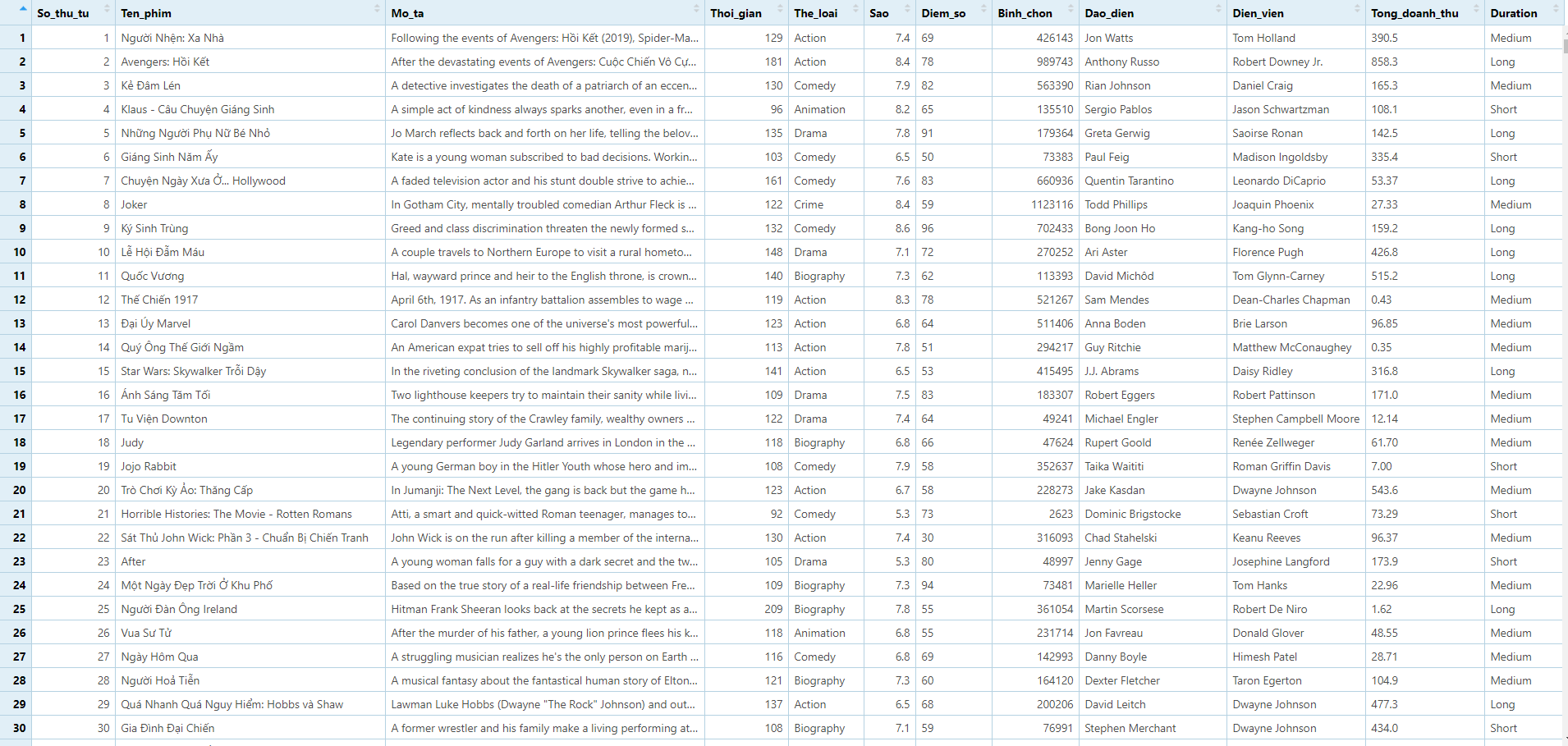
Thêm cột Duration với lượng tử mẫu tương ứng với các xác suất đã cho

runtime <= 108 : Short

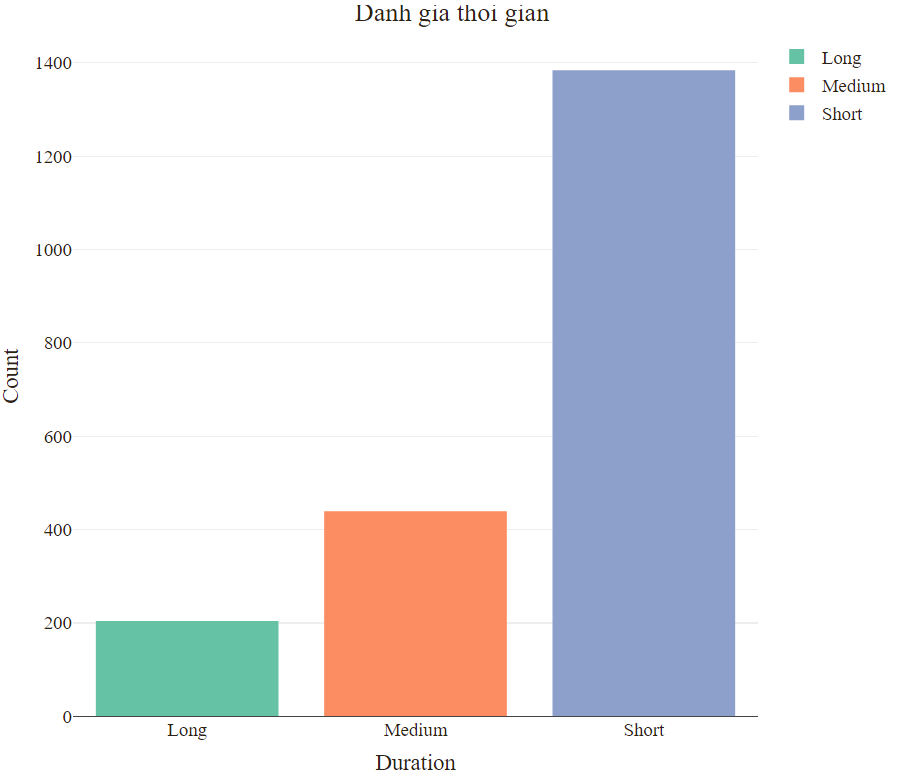
runtime <= 130: Medium

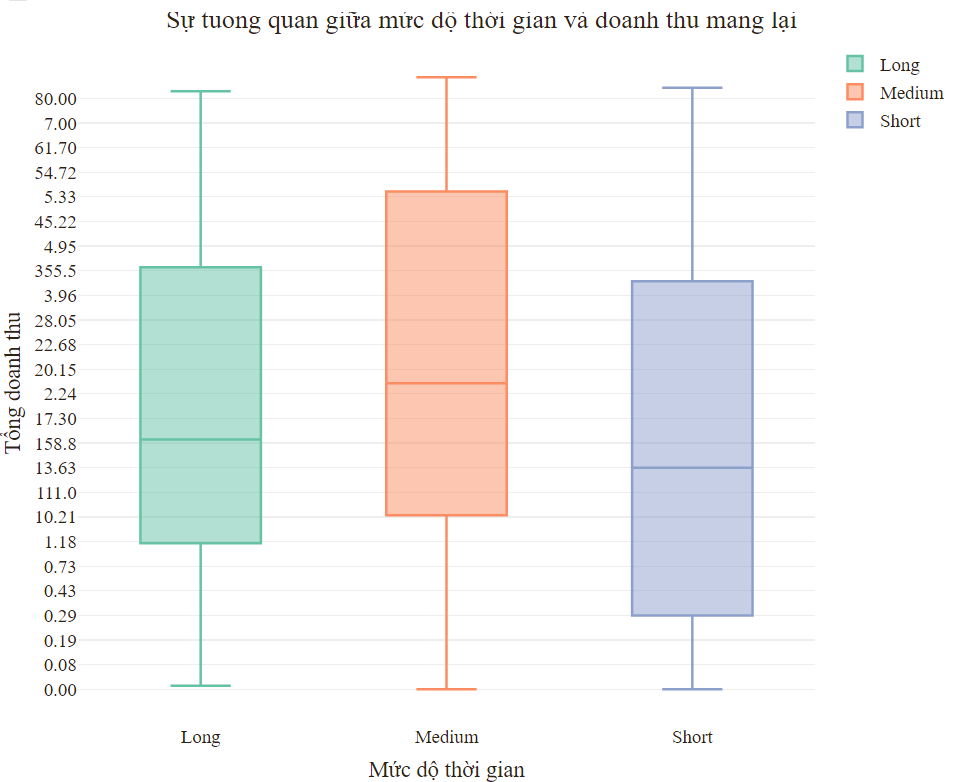
runtime >130 : Long

movies\_df sau khi thêm cột Duration



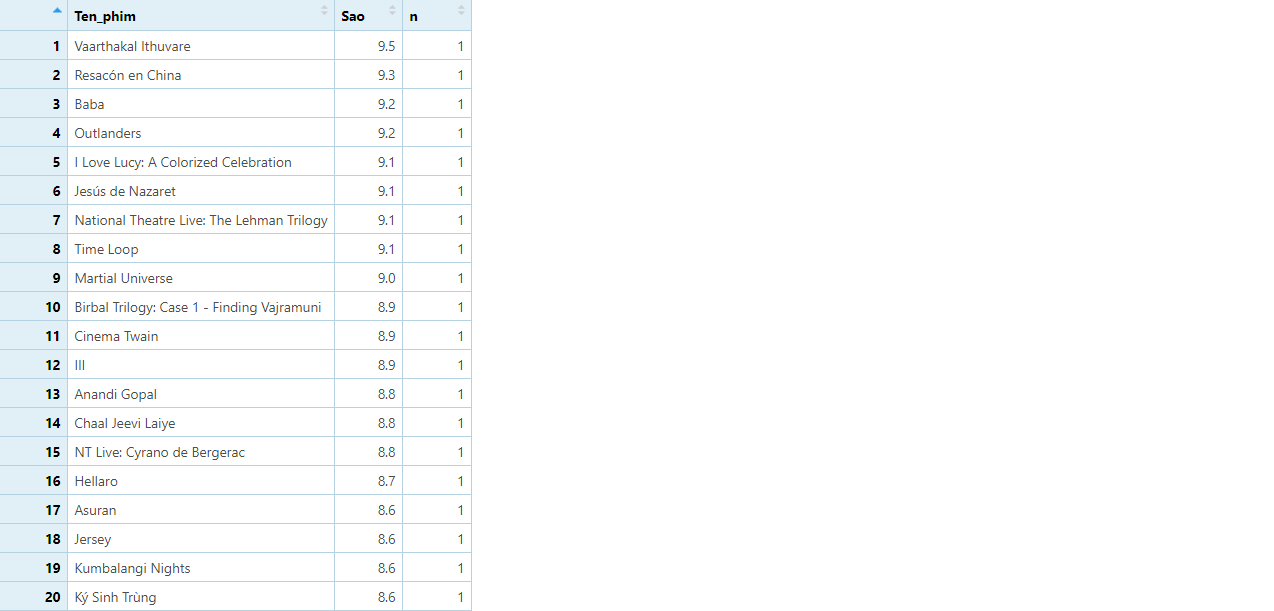
Biểu đồ histogram dễ hình dung hơn

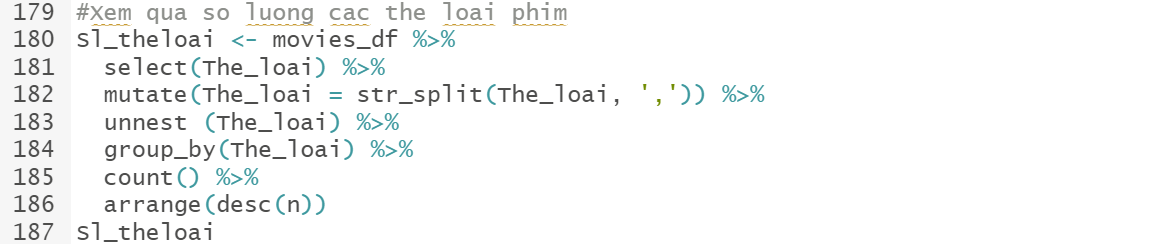


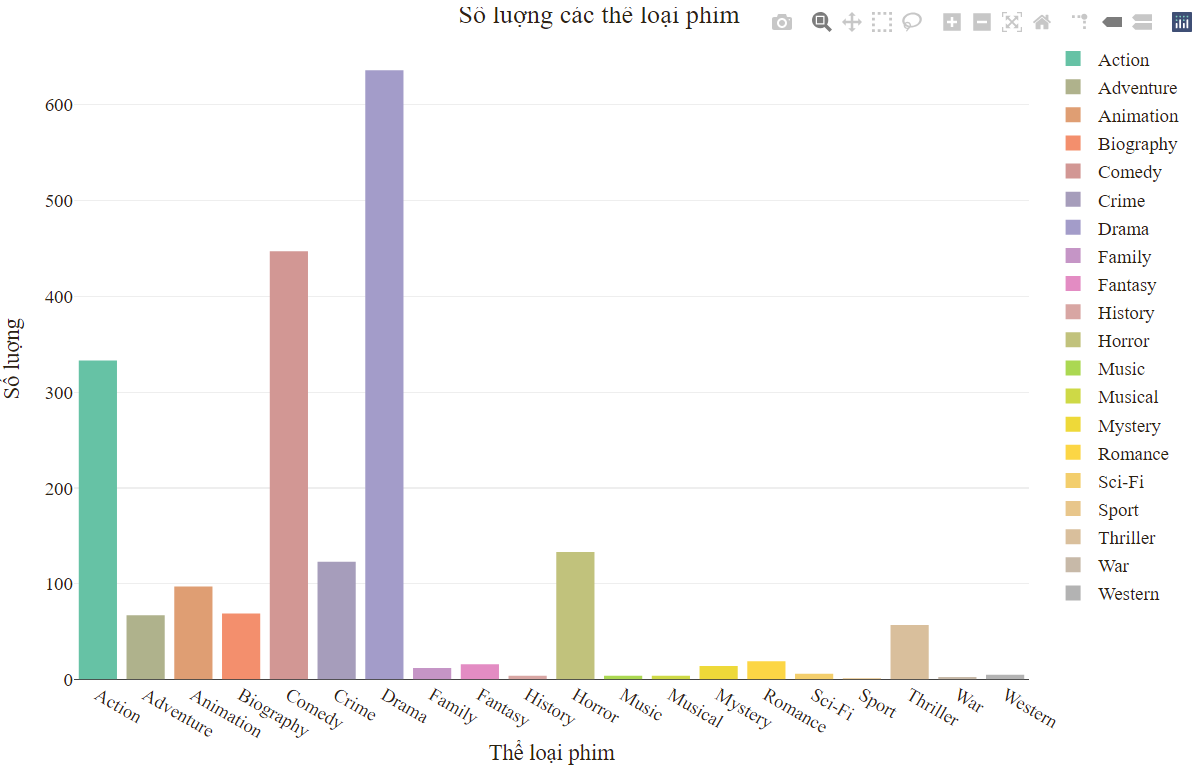


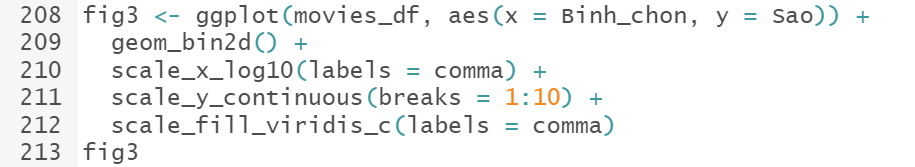
Kết quả cho thấy phim có thời gian ngắn đang chiếm phần lớn thị trường phim ảnh 2019

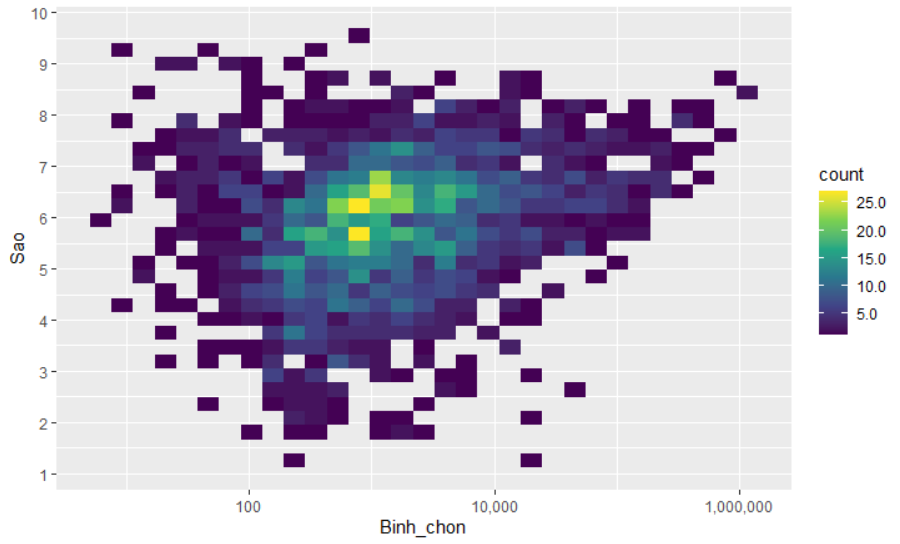
Top 20 phim được đánh giá nhiều sao nhất

 Xem qua số lượng các thể loại phim

  Biểu đồ số lượng các thể loại phim

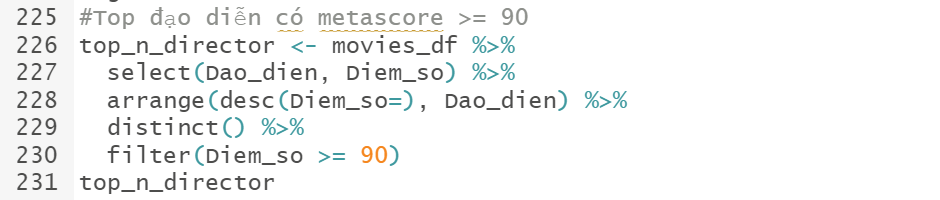
 Cho thấy phim Drama, Comedy, Action đang chiếm xu thế làm phim 2019

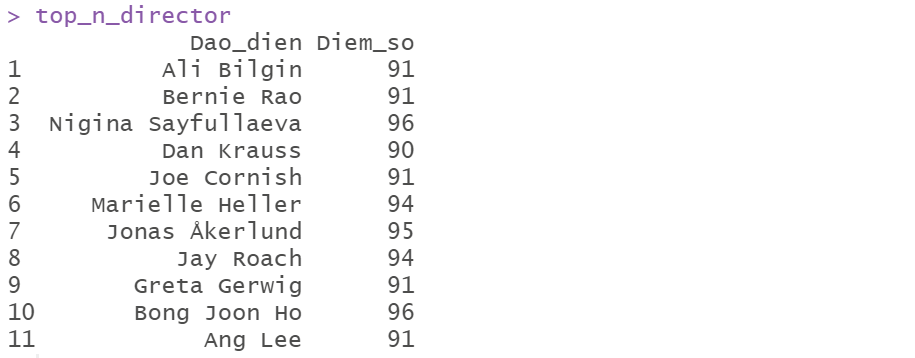




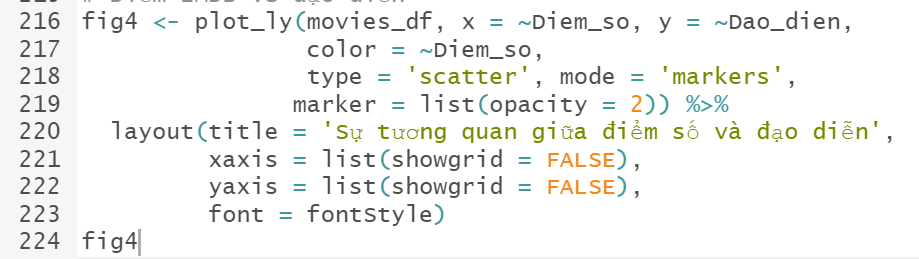
Biểu đồ cho thấy trung bình đánh giá sao chiếm từ 5-7 trong khoảng 200- 9000 bình chọn

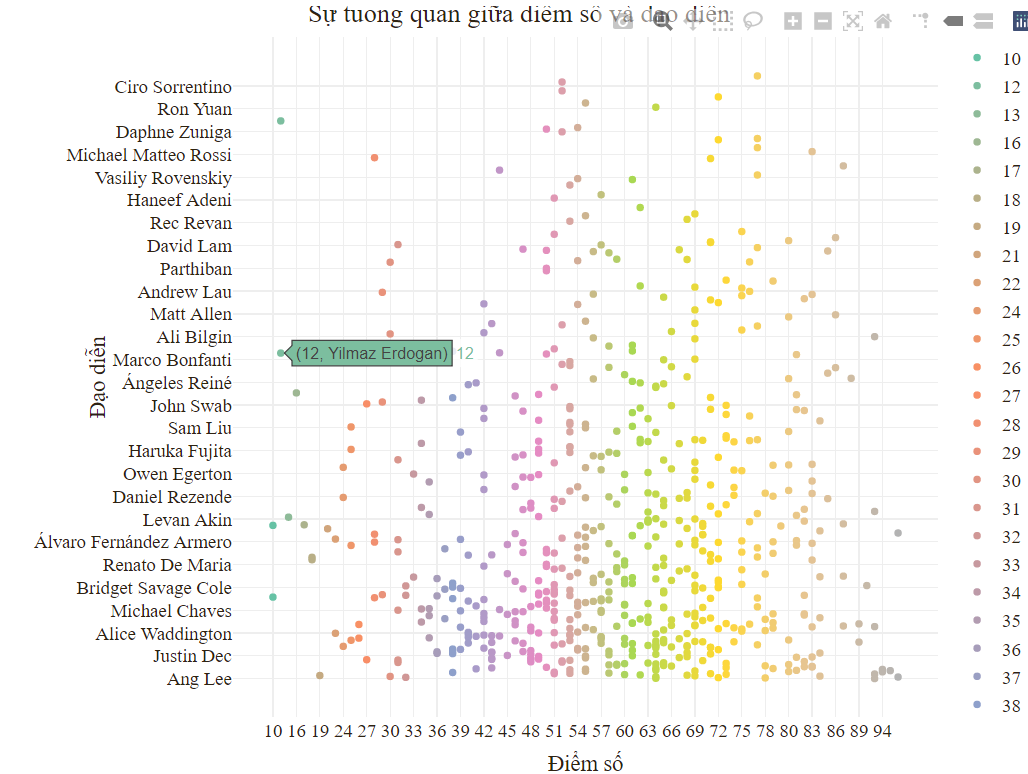
Code sự tương quan giữa điểm số phim và các đạo diễn



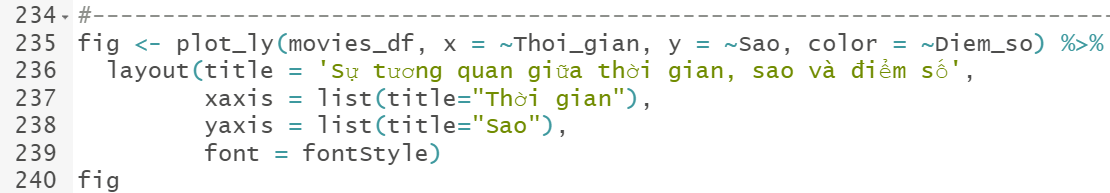


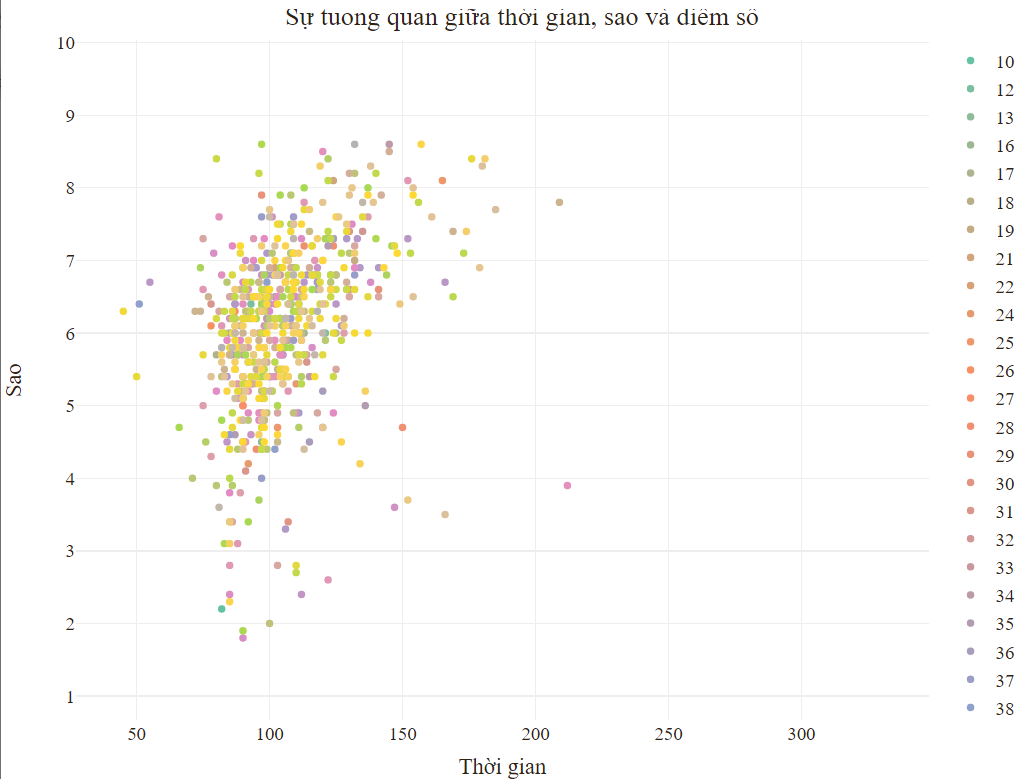
Chúng ta có cái nhìn cụ thể về nhiều đạo diễn qua biểu đồ như sau





Sự tương quan giữa thời gian, sao và điểm số

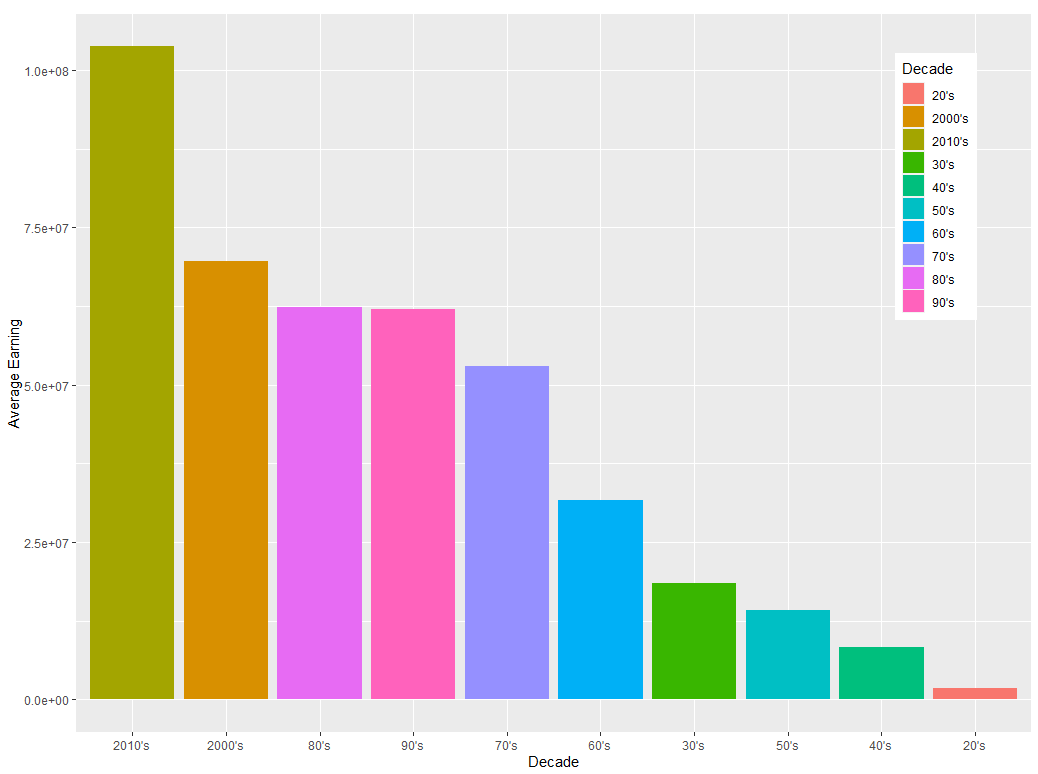




Phim có điểm từ 60-80 chiếm đa số, với số lượng lớn nằng trong khoảng 5-7 sao

* 1. Phân tích số liệu

Dựa vào dữ liệu từ kaggle. Phân tích đưa ra tiêu chuẩn cho việc sản xuất phim các năm tiếp theo

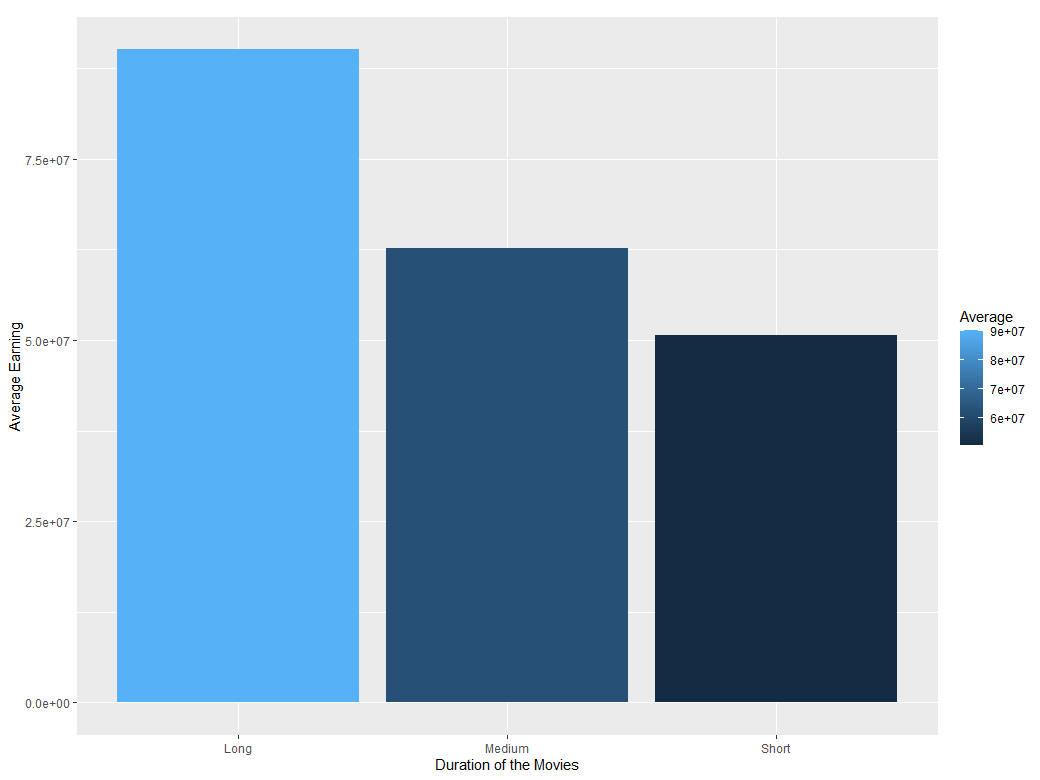


Như chúng ta có thể thấy thu nhập trung bình đã tăng lên ít nhiều trong nhiều thập kỷ, ngoại trừ một vài lần:

Thu nhập trung bình của các bộ phim ở độ tuổi 40 và 50 thấp hơn nhiều so với tuổi 30. Một lý do cho điều đó có thể là chiến tranh thế giới thứ 2 (1939-1945). Nó đã thay đổi tất cả các động lực về cách mọi người đang sống cuộc sống của họ. Và đúng ra, độ tuổi 40 và 50 chứng kiến ​​sự sụt giảm trong thu nhập từ Phim.

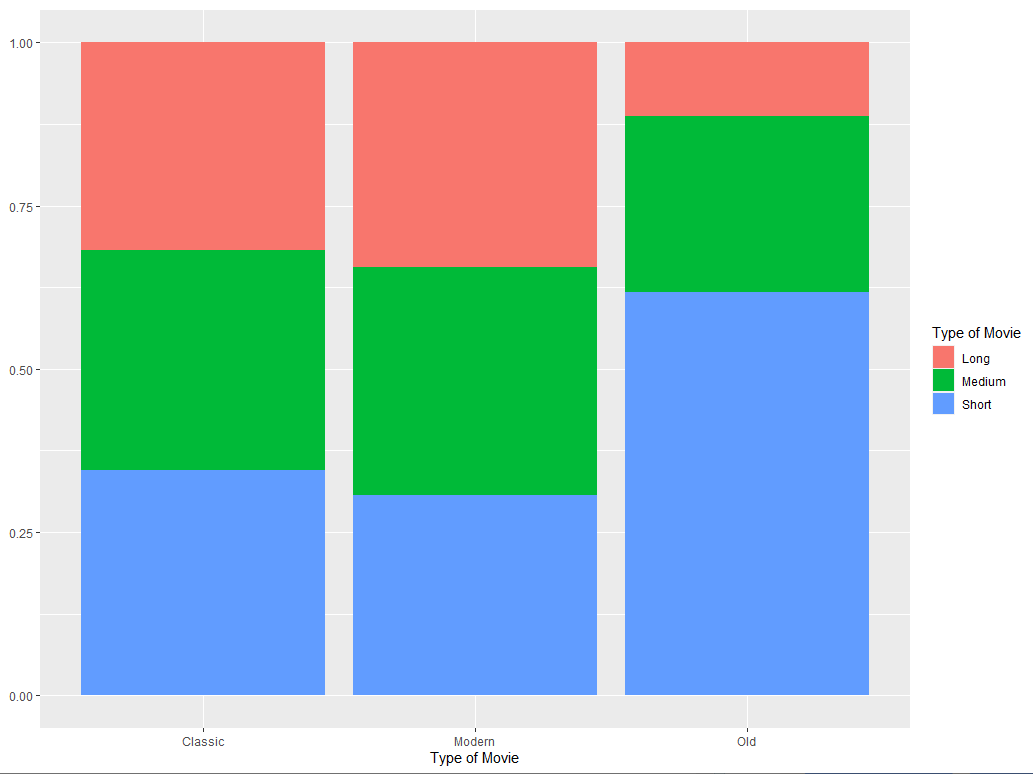
Thu nhập trung bình của những năm 80 cao hơn một chút so với những năm 90. Điều này có thể là do cuộc Đại suy thoái năm 1990

Số lượng phim ngắn nhiều hơn hai phim còn lại. Bây giờ ta xem qua biểu đồ Thu nhập trung bình theo Thời lượng Phim



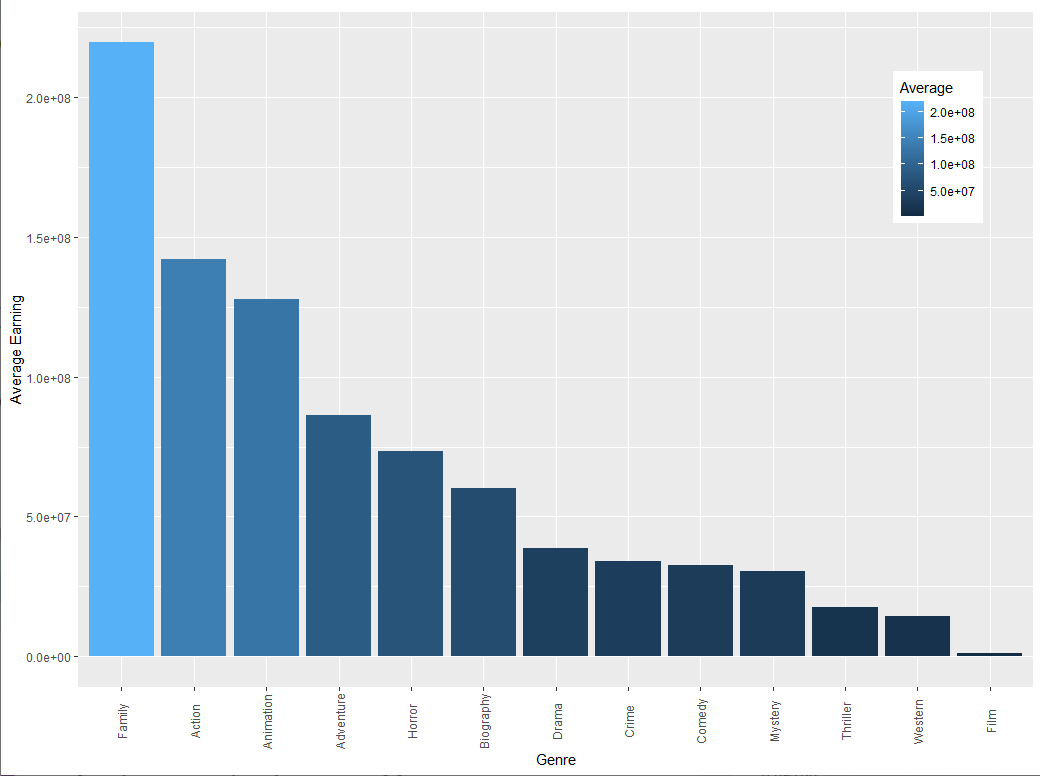
Mặc dù số lượng Phim ngắn là nhiều nhất, nhưng chúng ta có thể thấy ở đây rằng thu nhập từ Phim dài hơn nhiều hơn. Và bằng cách nào đó, thu nhập trung bình đã tăng lên khi thời lượng cũng tăng lên. Hiển thị mối tương quan giữa thời lượng và thu nhập.

Một điều cần xem xét ở đây là: Những bộ phim có thời lượng chạy khác nhau (dài, ngắn, trung bình) này có được phát hành thống nhất trong những năm qua không? hay chúng đã thay đổi trong nhiều thập kỷ khác nhau? Hãy cùng tìm hiểu.



Như chúng ta có thể thấy, Runtime không đồng nhất qua các thập kỷ khác nhau. Phim cũ có xu hướng ngắn hơn. Và số lượng phim Ngắn đã giảm ở 'độ tuổi' Hiện đại và Cổ điển. Vì vậy, điều đó giải thích tại sao phim ngắn hơn có thu nhập thấp nhất.

Một điều đáng chú ý khác là, không có sự khác biệt lớn giữa tỷ lệ phim Dài + Trung bình so với phim Hiện đại và Cổ điển. Nhưng chúng ta có thể đoán từ biểu đồ Thập kỷ rằng thu nhập từ phim Hiện đại sẽ nhiều hơn thu nhập kinh điển.



Như chúng ta có thể thấy, Thu nhập trung bình của Gia đình cao hơn nhiều so với những người khác.

Source code: <https://github.com/tandat-1305/R-/blob/main/Doan_2022_HKI/Code_R.R>

Source code: <https://github.com/tandat-1305/R-/blob/main/Doan_2022_HKI/Code_kg_R.R>