

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:**

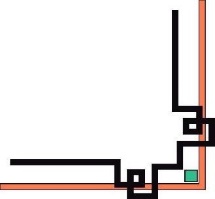
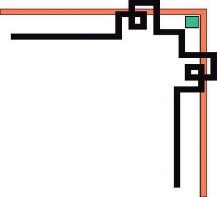
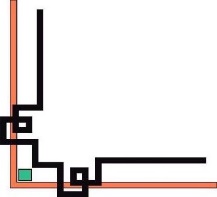
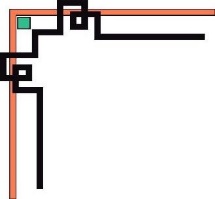
Phân tích dữ liệu R

***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 Tiến Anh Quân

###### 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

Báo cáo đồ án cuối kỳ

1. ***Mô tả về dữ liệu sử dụng crawl lấy về***
2. Số lượng dòng dữ liệu
3. Danh sách các trường dữ liệu (tên cột), kiểu dữ liệu (số, văn bản…)

***II. Phân tích thống kê mô tả số liệu (sử dụng R để tính toán và vẽ biểu đồ)***

1. Tính các chỉ số thống kê mô tả như mean, mode, median
2. Vẽ các biểu đồ để trực quan hóa dữ liệu
3. Một số nhận xét về số liệu

***III.Các tài liệu và trang web đã kham khảo***

***IV.Kiểm tra đạo văn***

**LỜI CẢM ƠN**

Tôi xin bày tỏ lời cảm ơn của mình đến thây HỒ QUỐC DŨNG, thuộc bộ ngôn ngữ lập trình R Khoa Kỹ Thuật và Công Nghệ, Đại học Huế.Trong quá trình thực hiện đồ án , thầy đã nhiệt tình giúp đỡ,giải đáp các thắc mặc tạo động lực giúp tôi hoàn thành đồ án này.

Báo cáo đồ án cuối kỳ

1. Mô tả về dữ liệu sử dụng crawl lấy về

Dữ liệu crawl

1. Số lượng dòng dữ liệu

Dữ liệu bao gồm 6 cột, 31 dòng.

2 . Danh sách các dữ liệu tên ,cột

|  |  |
| --- | --- |
| Tên Cột | Kiểu dữ liệu |
| STT | Num |
| Tên phim | Chr |
| Thời gian | Num |
| Thể loại | Factor |
| Điểm số | Chr |
| Bình chọn | Num |

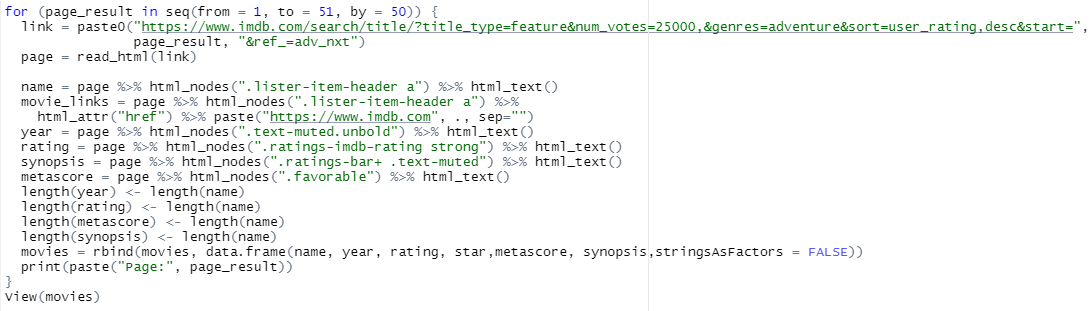
II. Phân tích thống kê mô tả số liệu (sử dụng R để tính toán và vẽ biểu đồ)

Phân tích thông kê mô tả dữ liệu

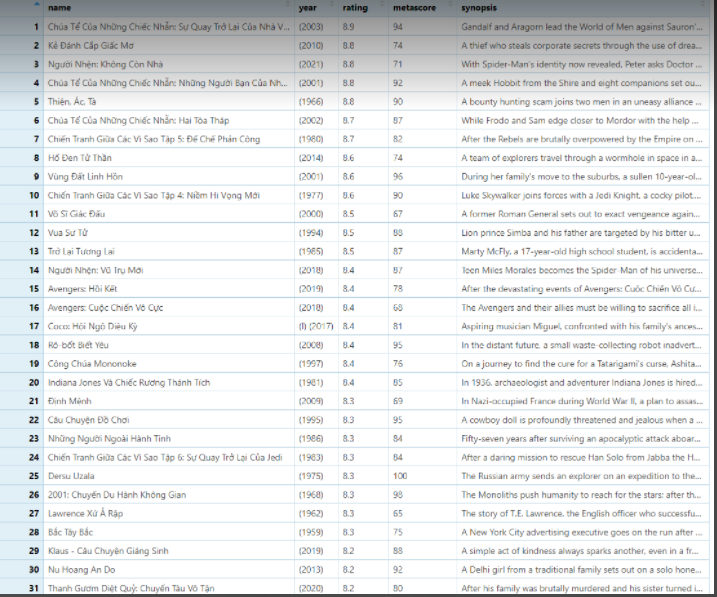
Các thư viện được sử dụngvà tạo một dattaframe



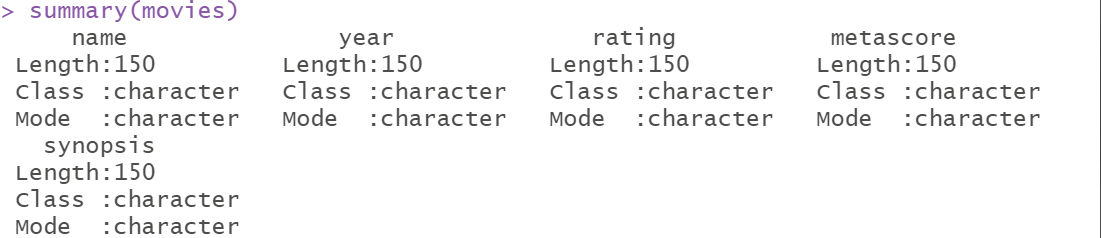
Code để crawl dữ liệu từ trang web về:



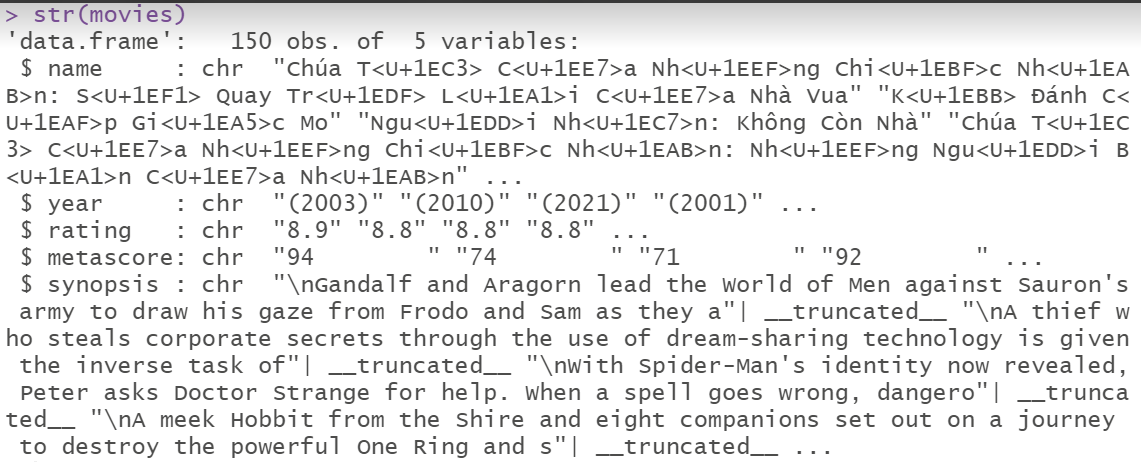
Bảng dữ liệu :



Tính summary cho dữ liệu



Kiểm tra định dạng các biến

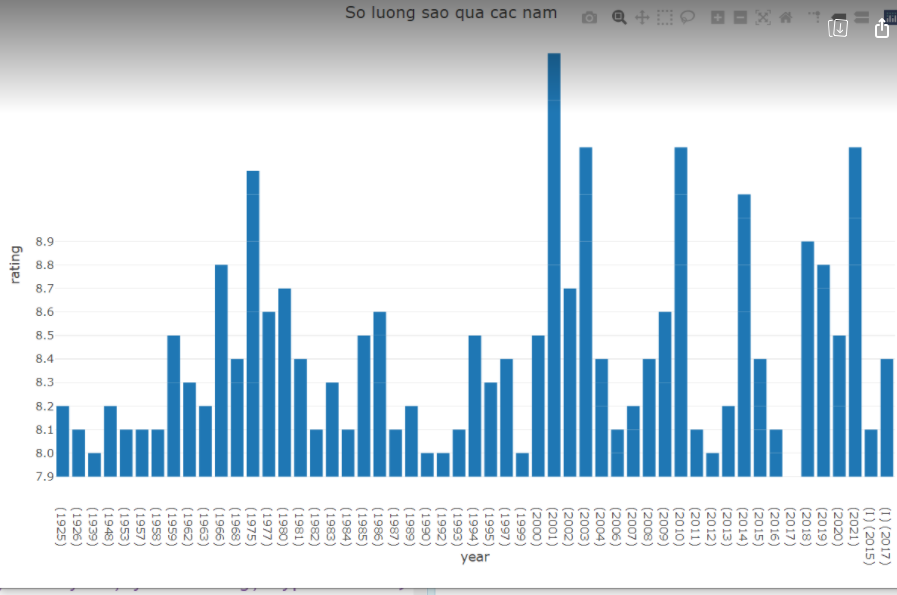


1. Vẽ các biểu đồ để trực quan hóa dữ liệu

Phân tích số liệu

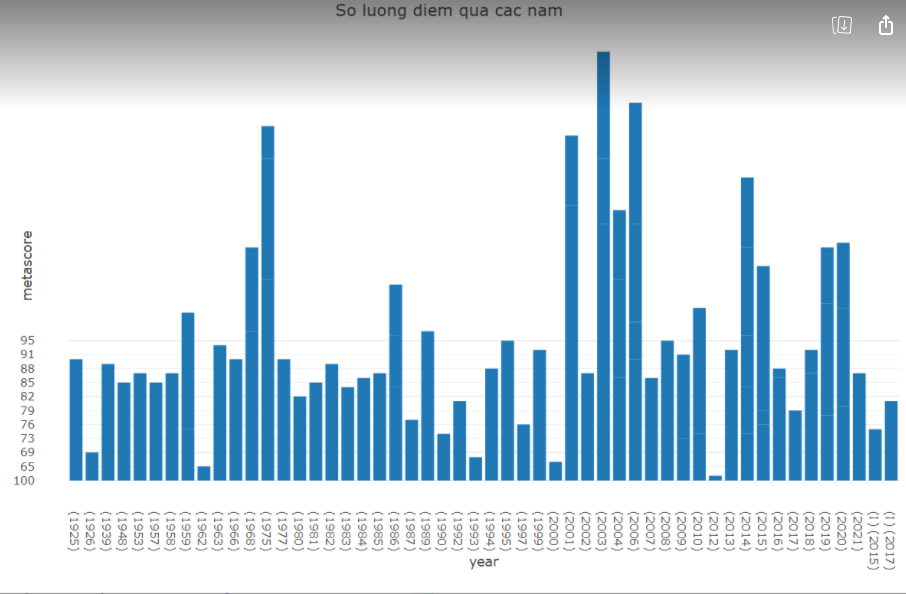
Dựa vào số liệu thì ta có các biểu đồ để biểu hiện 1 số so sánh :

-Bảng đanh giá số lượng sao trung của cả năm

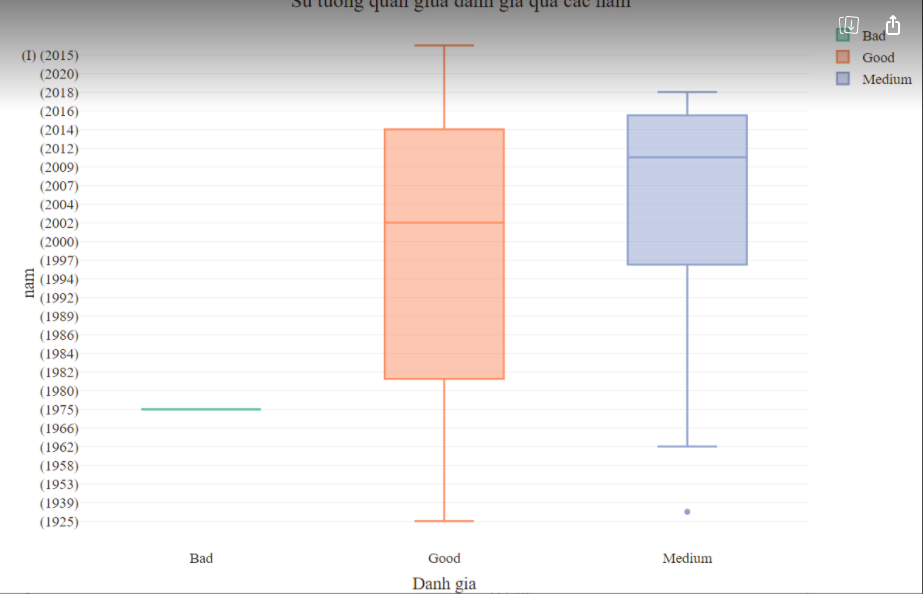


Như vậy ta thấy số lượng sao trung qua các năm không đồng đều điển hình vượt trội là năm 1975,2001, 2003,2010,2021.Các năm này số lượng sao cao đều trên 9.0. Còn có 1 số năm thì số sao chỉ là dười hoặc bằng 8.0 sao.

-Đánh giá tổng lựơng điểm số qua các năm



-Sự tương quan đánh giá phim qua các năm



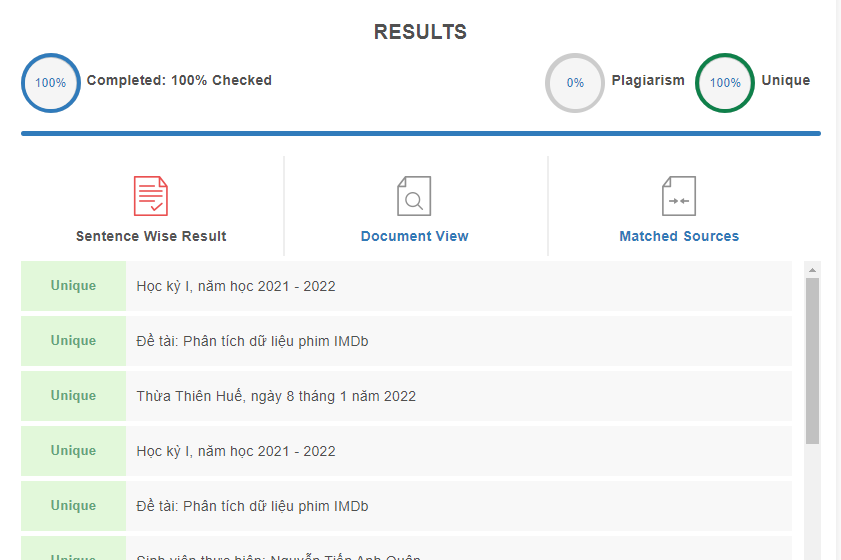
***III.Các tài liệu và trang web đã kham khảo***

***1.https://www.youtube.com/watch?v=28pyEDV9mMw***

***2.https://github.com/abhimotgi/dataslice/blob/master/R/Web%20Scraping%20with%20RVest%20Part%203.R/?fbclid=IwAR24QehXgGsqRLSzh78Qvbf9UZ3k-O\_jlgjscZNjqhYyrQVpE7EiGon4\_IY***

***3.https://www.kaggle.com/kanykeybegaliyeva/web-crawler-hw***

***IV.Kiểm tra đạo văn***

******

https://smallseotools.com/plagiarism-checker/