

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





# BÁO CÁO

## ĐỒ ÁN

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

**Phân tích dữ liệu với R**

***Đề tài:*** Phân tích dữ liệu Light Novel/Web Novel

…………………………………………………

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày …tháng…năm....



ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





# BÁO CÁO

## ĐỒ ÁN

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

**Phân tích dữ liệu với R**

***Đề tài:*** Phân tích dữ liệu Light Novel/Web Novel

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

###### Sinh viên thực hiện: Nguyễn Tất Duy Thành

###### Lớp: KHDL&TTNT

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày …tháng…năm.....

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 1](#_Toc92574677)

[ĐỒ ÁN 1](#_Toc92574678)

[BÁO CÁO 2](#_Toc92574679)

[ĐỒ ÁN 2](#_Toc92574680)

[1. Cơ sở lý thuyết 6](#_Toc92574681)

[1.1 Thống kê là gì? 6](#_Toc92574682)

[A/ Định nghĩa 6](#_Toc92574683)

[B/ Tầm quan trọng của thống kê 8](#_Toc92574684)

[1.2 Thực hiện thống kê trong R như thế nào? 9](#_Toc92574685)

[A/ Giới thiệu về ngôn ngữ lập trình R 9](#_Toc92574686)

[B/ Giới thiệu về thư viện ggplot2 9](#_Toc92574687)

[2. Crawl dữ liệu 11](#_Toc92574688)

[3. Cơ sở dữ liệu và tài liệu tham khảo 12](#_Toc92574689)

[4. Mô tả về dữ liệu sử dụng crawl lấy về 13](#_Toc92574690)

[5. Phân tích thống kê mô tả số liệu (sử dụng R để tính toán và vẽ biểu đồ) 13](#_Toc92574691)

[a. Phân tích thống kê mô tả của số liệu: 13](#_Toc92574692)

[b. Vẽ biểu đồ 16](#_Toc92574693)

[6. Thống kê suy diễn 20](#_Toc92574694)

[A/ Kiểm định T-test 20](#_Toc92574695)

[B/ Kiểm định Chi2 21](#_Toc92574696)

[B1. Dùng kiểm định chi2 để nghiên cứu ảnh hưởng của số lượt xem đến số chữ 21](#_Toc92574697)

[B.2/ Dùng kiểm định Chi2 để nghiên cứu ảnh hưởng của lượt theo dõi đến đánh giá 23](#_Toc92574698)

[B3/ Dùng kiểm định Chi2 để nghiên cứu ảnh hưởng của số chữ đến lượt theo dõi 25](#_Toc92574699)

[C/ Kiểm định kruskal-wallis 25](#_Toc92574700)

[7. Tài liệu tham khảo 28](#_Toc92574701)

# 1. Cơ sở lý thuyết

## 1.1 Thống kê là gì?

### **A/ Định nghĩa**

- Thống kê là một nhánh của toán học ứng dụng liên quan đến việc thu thập, mô tả, phân tích và suy luận các kết luận từ dữ liệu định lượng. Các lý thuyết toán học đằng sau thống kê chủ yếu dựa vào phép tính vi phân và tích phân, đại số tuyến tính và lý thuyết xác suất.

- Trên thực tế, thống kê là ý tưởng chúng ta có thể tìm hiểu về các tính chất của bộ lớn của các đối tượng hoặc các sự kiện (một quần thể) bằng cách nghiên cứu các đặc điểm của một số nhỏ các đối tượng hoặc các sự kiện (tương tự mẫu ). Bởi vì trong nhiều trường hợp việc thu thập dữ liệu toàn diện về toàn bộ dân số là quá tốn kém, khó khăn hoặc không thể thực hiện được, thống kê bắt đầu với một mẫu có thể được quan sát một cách thuận tiện hoặc có thể chi trả được.

- Các nhà thống kê, những người làm công tác thống kê, đặc biệt quan tâm đến việc xác định cách đưa ra kết luận đáng tin cậy về các nhóm lớn và các hiện tượng chung từ các đặc điểm quan sát được của các mẫu nhỏ chỉ đại diện cho một phần nhỏ của nhóm lớn hoặc một số trường hợp hạn chế của một hiện tượng chung.

- Thống kê được sử dụng hầu như trong tất cả các ngành khoa học như khoa học vật lý và xã hội, cũng như trong kinh doanh, nhân văn, chính phủ và sản xuất.

- Lợi ích của thống kê:

* Thống kê sử dụng bằng chứng số để rút ra kết luận hợp lệ
* Các nhà thống kê biết cách tránh những sai lầm thường gặp

- Hai lĩnh vực chính của thống kê là:

* Thống kê mô tả.
* Thống kê suy diễn.

**a. Thống kê mô tả**

- Mô tả thống kê là hệ số mô tả ngắn gọn rằng tóm tắt một tập dữ liệu nhất định, đó có thể là một đại diện của toàn bộ hoặc một mẫu của một dân số. Thống kê mô tả được chia thành các thước đo về xu hướng trung tâm và các thước đo về độ biến thiên (lan truyền). Các phép đo xu hướng trung tâm bao gồm giá trị trung bình, giá trị trung bình và phương thức, trong khi các phép đo độ biến thiên bao gồm độ lệch, phương sai, các biến số tối thiểu và tối đa, [kurtosis](https://www.investopedia.com/terms/k/kurtosis.asp) và [độ lệch](https://www.investopedia.com/terms/s/skewness.asp) .

- Thống kê mô tả được sử dụng để mô tả những đặc tính cơ bản của [dữ liệu](https://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u) thu thập được từ nghiên cứu thực nghiệm qua các cách thức khác nhau. Thống kê mô tả và [thống kê suy luận](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_suy_lu%E1%BA%ADn) cùng cung cấp những tóm tắt đơn giản về [mẫu](https://vi.wikipedia.org/wiki/M%E1%BA%ABu_(th%E1%BB%91ng_k%C3%AA)) và các thước đo. Cùng với phân tích đồ họa đơn giản, chúng tạo ra nền tảng của mọi phân tích định lượng về số liệu. Để hiểu được các hiện tượng và ra quyết định đúng đắn, cần nắm được các phương pháp cơ bản của mô tả dữ liệu. Có rất nhiều kỹ thuật hay được sử dụng. Có thể phân loại các kỹ thuật này như sau:

* Biểu diễn dữ liệu bằng đồ họa trong đó các đồ thị mô tả dữ liệu hoặc giúp so sánh dữ liệu;
* Biểu diễn dữ liệu thành các bảng số liệu tóm tắt về dữ liệu;
* Thống kê tóm tắt (dưới dạng các giá trị thống kê đơn nhất) mô tả dữ liệu.

Khi tạo các trị thống kê mô tả, người ta có thể nhằm 2 mục tiêu:

* Chọn một trị thống kê để chỉ ra những đơn vị có vẻ giống nhau thực ra có thể khác nhau thế nào. Các giáo trình thống kê gọi một giải pháp đáp ứng mục tiêu này là thước đo khuynh hướng trung tâm.
* Chọn một trị thống kê khác cho thấy các đơn vị khác nhau thế nào. Loại trị thống kê này thường được gọi là một thước đo phân tán thống kê.

Khi tóm tắt một lượng như độ dài, cân nặng hay tuổi tác, nói chung người ta hay dùng các trị thống kê như số trung bình cộng, trung vị; hay trong trường hợp một phân bố đơn mốt (mode - số trung phương), người ta thường dùng mốt. Đôi khi, người ta chọn lựa những giá trị đặc thù từ hàm phân bố tích lũy gọi là các tứ phân vị.

Các thước đo chung nhất về mức độ phân tán của dữ liệu lượng là [phương sai](https://vi.wikipedia.org/wiki/Ph%C6%B0%C6%A1ng_sai), giá trị căn bậc 2 của nó, tức là độ lệch chuẩn; khoảng; khoảng cách giữa các tứ phân vị; và độ lệch bình quân tuyệt đối.

Khi thực hiện một trình diễn đồ họa để tóm tắt một bộ dữ liệu, cũng có thể áp dụng cả hai mục tiêu nói trên. Một ví dụ đơn giản về kỹ thuật đồ họa là [đồ thị phân bố](https://vi.wikipedia.org/w/index.php?title=%C4%90%E1%BB%93_th%E1%BB%8B_ph%C3%A2n_b%E1%BB%91&action=edit&redlink=1), thứ đồ thị phơi bày cả khuynh hướng trung tâm lẫn độ [phân tán thống kê](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_t%C3%A1n_th%E1%BB%91ng_k%C3%AA&action=edit&redlink=1).

**b. Thống kê suy diễn**

- Thống kê suy diễn là công cụ mà nhà thống kê sử dụng để rút ra kết luận về các đặc điểm của một quần thể từ các đặc điểm của mẫu và quyết định mức độ chắc chắn của chúng về độ tin cậy của các kết luận đó.

- Dựa trên kích thước mẫu và sự phân bố của dữ liệu mẫu, các nhà thống kê có thể tính toán xác suất để các thống kê đo lường xu hướng trung tâm, độ biến thiên, phân bố và mối quan hệ giữa các đặc điểm trong mẫu dữ liệu, cung cấp bức tranh chính xác về các thông số tương ứng của toàn bộ tổng thể từ đó mẫu được rút ra.

- Thống kê suy luận được sử dụng để khái quát hóa về các nhóm lớn, chẳng hạn như ước tính nhu cầu trung bình đối với một sản phẩm bằng cách khảo sát mẫu thói quen mua hàng của người tiêu dùng hoặc để cố gắng dự đoán các sự kiện trong tương lai, chẳng hạn như dự đoán lợi nhuận trong tương lai của một loại chứng khoán hoặc tài sản về lợi nhuận trong một khoảng thời gian mẫu.

### **B/ Tầm quan trọng của thống kê**

Ngày nay, thống kê đóng vai trò quan trọng trong nhiều lĩnh vực:

- Học qua máy và khai thác dữ liệu: có hai ứng dụng cho học qua máy móc và khai thác dữ liệu: quản lý dữ liệu và phân tích dữ liệu. Các công cụ thống kê cần thiết cho việc phân tích dữ liệu.

- Thống kê trong xã hội học: thống kê được áp dụng cho một loạt các môn học, bao gồm cả khoa học tự nhiên và xã hội, chính trị và kinh doanh. Thống kê tư vấn có thể giúp các tổ chức và công ty không có chuyên môn trả lời những thắc mắc.

- Lĩnh vực chuyên môn: Các kỹ thuật thống kê được sử dụng trong một loạt các nghiên cứu khoa học và xã hội, bao gồm: ngành sinh học, tính toán sinh học, tính toán xã hội học, hệ thống sinh học, khoa học xã hội và nghiên cứu xã hội. Những ngành này bao gồm:

* Khoa học tính toán bảo hiểm
* Ứng dụng thông tin kinh tế
* Thiên văn học
* Sinh học
* Thống kê kinh doanh
* Hóa học (phân tích dữ liệu hóa học)
* Khai thác dữ liệu (áp dụng thống kê và nhận dạng mẫu để khám phá tri thức từ dữ liệu)
* Dân số học
* Kinh tế học (phân tích thống kê các số liệu kinh tế)
* Thống kê năng lượng
* Thống kê kỹ thuật
* Khoa học nghiên cứu bệnh dịch (phân tích thống kê của bệnh)
* Địa lý và hệ thống thông tin địa lý, đặc biệt trong phân tích không gian
* Xử lý hình ảnh
* Thống kê y tế
* Thống kê về hành vi, tâm lý
* Độ bền cơ khí
* Thống kê xã hội

- Thống kê áp dụng cho toán học hay nghệ thuật

## 1.2 Thực hiện thống kê trong R như thế nào?

### **A/ Giới thiệu về ngôn ngữ lập trình R**

R là một công cụ rất mạnh cho học máy, thống kê và phân tích dữ liệu. Nó là một ngôn ngữ lập trình. Ngôn ngữ R là một platform-independent do đó chúng ta có thể sử dụng nó cho bất kỳ hệ điều hành nào. Việc cài đặt R cũng miễn phì vì thế chúng ta có thể sử dụng mà không cần phải mua bản quyền.

Ngôn ngữ R là mã nguồn mở nên ai trong chúng ta cũng có thể phân tích source code để hiểu được chính xác cách R vận hành. Bất kỳ ai cũng có thể thêm tính năng và fix bug mà không cần chờ nhà phát hành ra bản vá. Đồng thời, R có thể tích hợp được với ngôn ngữ khác (C,C++). Nó cũng cho phép chúng ta tương tác với nhiều nguồn dữ liệu và các gói thống kê (SAS, SPSS). R có một cộng đồng phát triển mạnh mẽ.

### **B/ Giới thiệu về thư viện ggplot2**

Ggplot2 là một package hỗ trợ visualization rất mạnh trong R. Dựa trên package này ta có thể vẽ được các đồ thị dạng barchart, line, plot, density, candle chart,pie,… và rất nhiều các đồ thị khác.Ngoài ra ggplot2 còn cho phép người dùng tùy chỉnh màu sắc, kích cỡ, theme, title, … để đồ thị được đẹp hơn. Cấu trúc của ggplot2 được chia rõ ràng làm 2 phần chính.

Ggplot (): phần này qui định đồ thị sẽ sử dụng data nào làm đầu vào. Lưu ý data phải có dạng data.frame. Dạng vector sẽ không được support.

geom\_(aes(x,y)): Phần này qui định kiểu đồ thị và các trục tọa độ từ dữ liệu đầu vào. Nếu chỉ có ggplot() mà không thêm geom\_() thì chúng ta chỉ nhận được background mà không có đồ thị mặc dù data đã được khai báo. Trong geom\_() chúng ta phải khai báo thêm trục tọa độ vào các arguments x và y của aes() chẳng hạn như geom\_point(aes(x=bienx,y=bieny)).

Các kiểu đồ thị chính: geom\_line(): biển diễn line geom\_point(): biểu diễn point geom\_chart(): biểu diễn chart geom\_density(): biểu diễn dưới dạng density geom\_vertical(): thêm trục vertical geom\_abline(): thêm trục horizontal, vertical và diagonal geom\_qq(): quantile and quantile plot geom\_contour(): vẽ các đường đồng mức 2d và 3d geom\_label() geom\_text(): hiển thị text geom\_raster(), geom\_tile(), geom\_rect(): biểu diễn dạng màu sắc mật độ.

**1.3 Thực hiện thống kê trong Python như thế nào?**

**A/ Giới thiệu về ngôn ngữ lập trình Python**

- Python là một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng, do Guido van Rossum tạo ra và lần đầu ra mắt vào năm 1991.

- Python được thiết với ưu điểm là dễ đọc, dễ học và dễ nhớ. Được phát triển theo phương châm là cấu trúc rõ ràng, hình thức sáng sủa, thuận tiện cho người dùng. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu.

- Ban đầu, Python được phát triển để chạy trên nền Unix. Nhưng rồi theo thời gian, Python dần mở rộng sang mọi hệ điều hành từ MS-DOS đến Mac OS, OS/2, Windows, Linux và các hệ điều hành khác thuộc họ Unix.

**B/ Giới thiệu về thư viện Pandas**

- Pandas là một Thư viện Python mã nguồn mở cung cấp công cụ phân tích và thao tác dữ liệu hiệu suất cao bằng cách sử dụng các cấu trúc dữ liệu mạnh mẽ của nó. Cái tên Pandas có nguồn gốc từ từ Panel Data - một Kinh tế lượng từ dữ liệu Đa chiều.

- Năm 2008, nhà phát triển Wes McKinney bắt đầu phát triển gấu trúc khi cần công cụ phân tích dữ liệu hiệu suất cao, linh hoạt.

- Trước Pandas, Python chủ yếu được sử dụng để trộn và chuẩn bị dữ liệu. Nó có rất ít đóng góp vào việc phân tích dữ liệu. Gấu trúc đã giải quyết vấn đề này. Sử dụng Pandas, chúng ta có thể thực hiện năm bước điển hình trong quá trình xử lý và phân tích dữ liệu, bất kể nguồn gốc của dữ liệu - tải, chuẩn bị, thao tác, lập mô hình và phân tích.

- Python with Pandas được sử dụng trong nhiều lĩnh vực bao gồm các lĩnh vực học thuật và thương mại bao gồm tài chính, kinh tế, thống kê, phân tích,…

- Các tính năng của Pandas:

* Đối tượng DataFrame nhanh và hiệu quả với lập chỉ mục mặc định và tùy chỉnh.
* Các công cụ để tải dữ liệu vào các đối tượng dữ liệu trong bộ nhớ từ các định dạng tệp khác nhau.
* Căn chỉnh dữ liệu và tích hợp xử lý dữ liệu bị thiếu.
* Định hình lại và xoay vòng các nhóm ngày.
* Cắt dựa trên nhãn, lập chỉ mục và tập hợp con các tập dữ liệu lớn.
* Có thể xóa hoặc chèn các cột từ cấu trúc dữ liệu.
* Nhóm theo dữ liệu để tổng hợp và biến đổi.
* Kết hợp và kết hợp dữ liệu hiệu suất cao.
* Chức năng Chuỗi thời gian.

# 2. Crawl dữ liệu

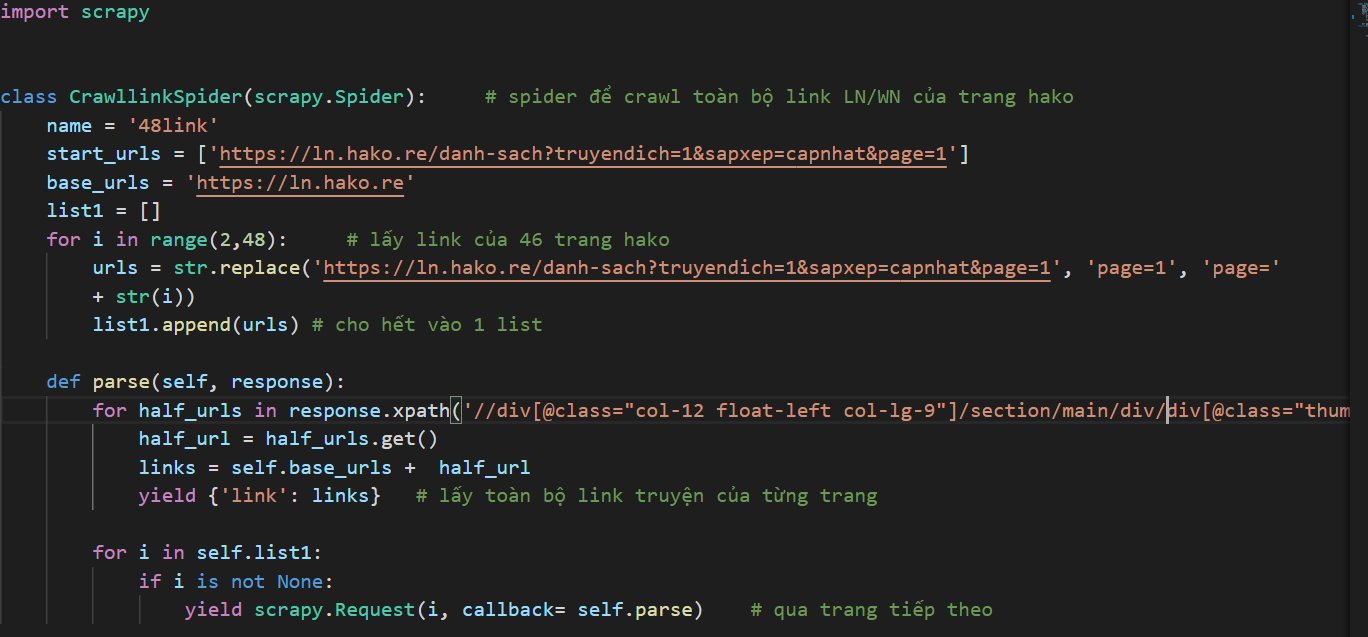
**2.1 Khái niệm**

Crawl (Spider hay là Bot) là một thuật ngữ mô tả quá trình thu thập dữ liệu trên web được gọi chung là Web Crawler. Phần mềm này được thiết kế để có thể duyệt website trên mạng World Wide Web, chúng tiến hành phân tích mã nguồn HTML, đọc dữ liệu và lọc ra theo yêu cầu của người sử dụng. Quá trình này sẽ dừng lại cho đến khi con bot duyệt hết tất cả các đường liên kết trang đầu cũng như các trang có liên quan.

**2.2 Crawl**

- Gồm 2 phần

+ Phần 1: chứa các link để crawl

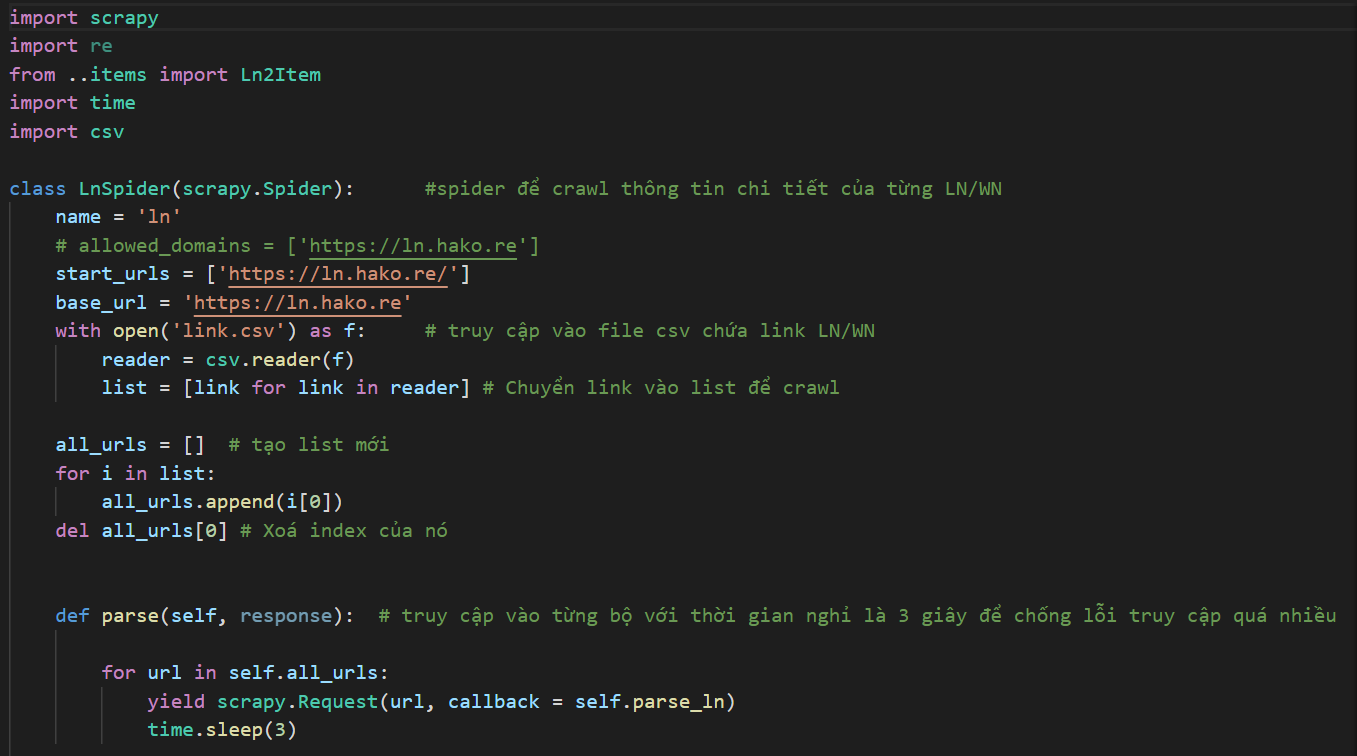


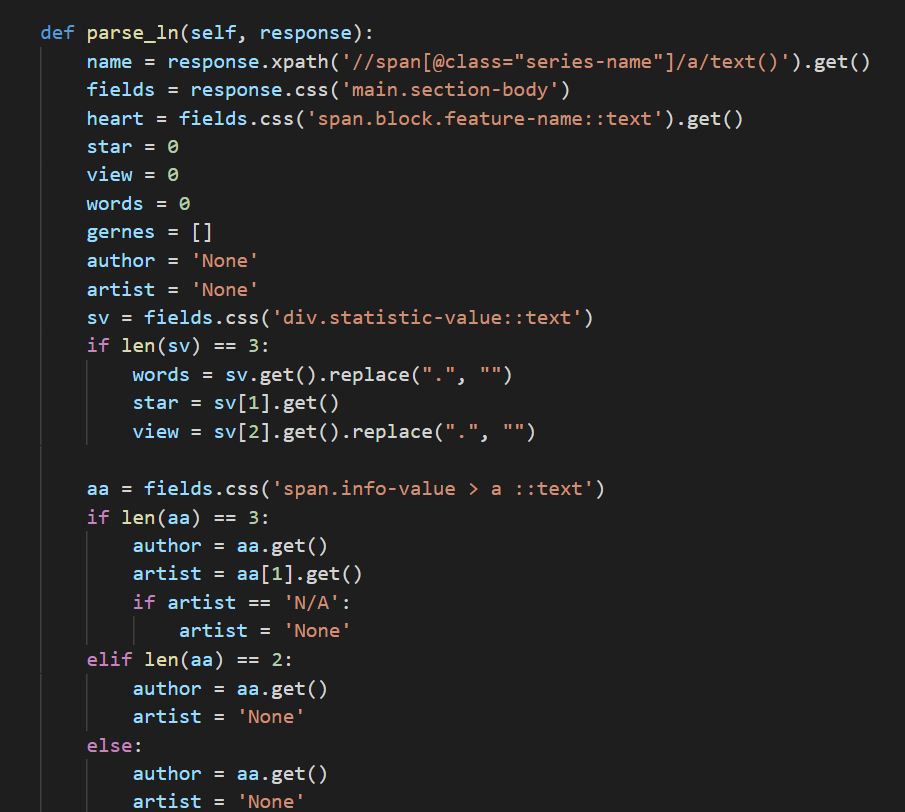
+ Phần 2: crawl dữ liệu của từng link

Trước khi scrapy và sau khi đưa ra ý tưởng, phân tích html, tôi chắc chắn rằng bạn đã định hình được dữ liệu bạn bao gồm những feature gì. Tệp “items.py” dùng để khai báo tên trường dữ liệu (cột dữ liệu). Với câu lệnh “tên\_trường\_dữ\_liệu = scrapy.Field()”. Một cách khác là sử dụng lệnh “yield” trong scrapy. Trong bài viết này, tôi sử dụng lệnh yield để lưu dữ liệu. Ở đây, tôi có các feature bao gồm:

* Artist: Tên hoạ sĩ
* Actor: Tên tác giả
* Name: Tên tác phẩm
* Star: đánh giá của tác phẩm
* View: Lượt xem của tác phẩm
* Heart: Lượt theo dõi của tác phẩm
* Gernes: Thể loại
* Words: Số lượng chữ

**Tiến hành crawl:**







# 3. Cơ sở dữ liệu và tài liệu tham khảo

- Cơ sở dữ liệu: Cổng Light Novel- Trang đọc Light Novel lớn nhất Việt Nam

Đường link: <https://ln.hako.re/>

- Nguồn tham khảo:

<https://www.statology.org/ggplot-pie-chart/>

<http://r-statistics.co/ggplot2-Tutorial-With-R.html>

<https://www.statology.org/best-ggplot2-themes/>

<https://rpubs.com/phamdinhkhanh/381826>

**4. Mô tả về dữ liệu sử dụng crawl lấy về**

**a. Số lượng dòng và cột dữ liệu:**

- 8 cột x 1934 hàng

**b. Danh sách các trường dữ liệu (tên cột), kiểu dữ liệu (số, văn bản…)**

- Các trường dữ liệu: artist, author, gernes, heart, name, star, view, words.

- Kiểu dữ liệu:

+ Kiểu văn bản: artist, author, gernes, name

+ Kiểu số: heart, star, view, words

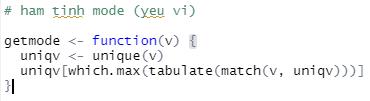


*Hình 1: 20 dòng đầu tiên của dữ liệu.*

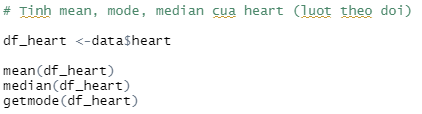
**5. Phân tích thống kê mô tả số liệu (sử dụng R để tính toán và vẽ biểu đồ)**

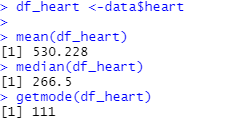
**a. Phân tích thống kê mô tả của số liệu:**

\* Tạo một hàm để tính mode (yếu vị):

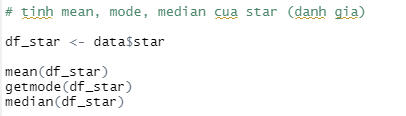


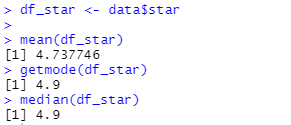
- Tính mean, mode, median của Heart (lượt theo dõi)



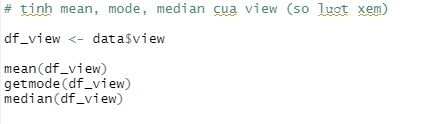


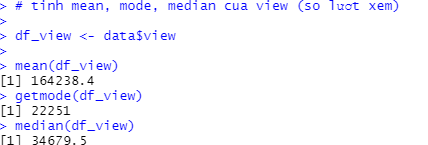
**- Tính mean, mode, median của star (đánh giá)**



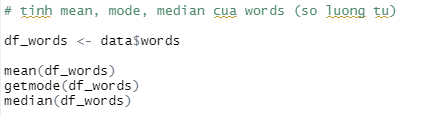


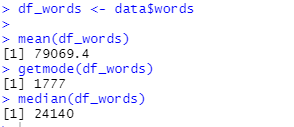
- Tính mean, mode, median của view (số lượt xem):



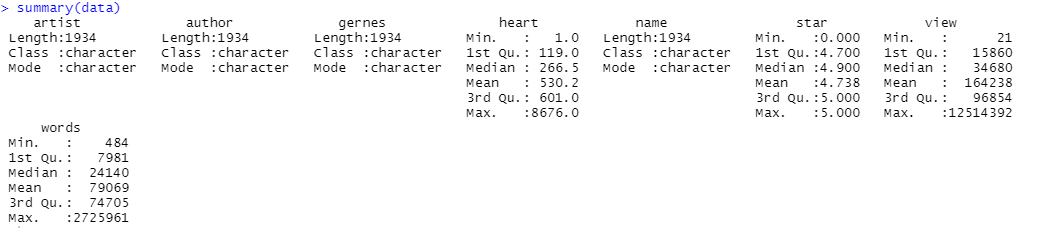


**- Tính mean, mode, median của words (số lượng chữ)**



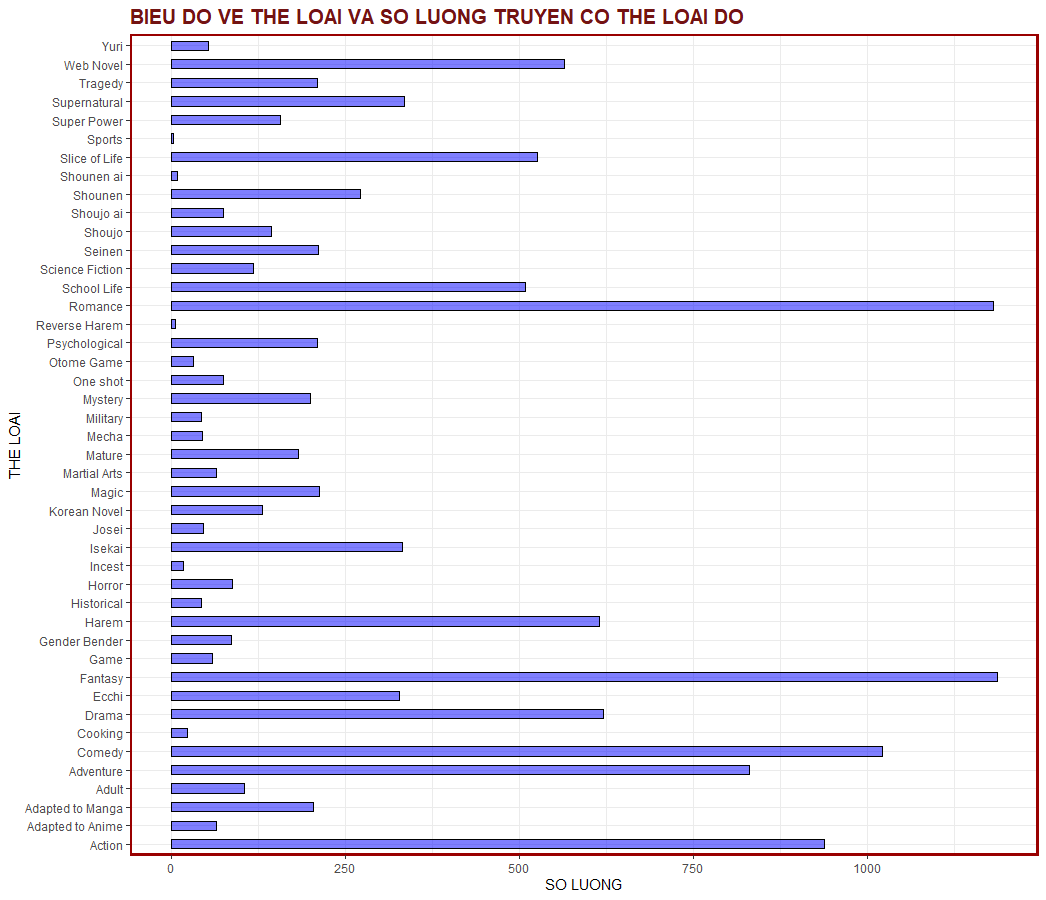


\* Tổng quan về dữ liệu (sử dụng hàm summary)



## b. Vẽ biểu đồ

**- Plot 1:**

****

Nhận xét:

- Biểu đồ dạng cột

- Biểu đồ biểu diễn số thể loại và số LN/WN có chứa thể loại đấy

+ X là cột số lượng

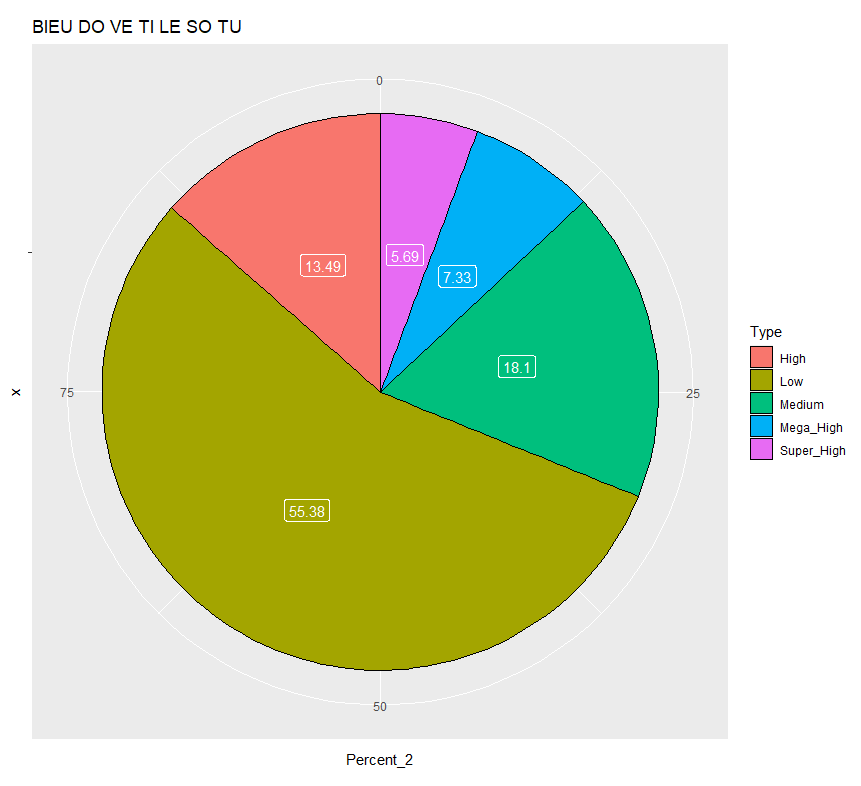
+ Y là cột thể loại

- Nhận xét về thể loại: phong phú, đa dạng với nhiều thể loại khác nhau. Trong đó:

+ Thể loại chiếm nhiều nhất là Fantasy (Phiêu lưu)

+ Thể loại chiếm ít nhất là: Sport (Thể thao)

**-Plot 2:**



Nhận xét:

- Biểu đồ tròn

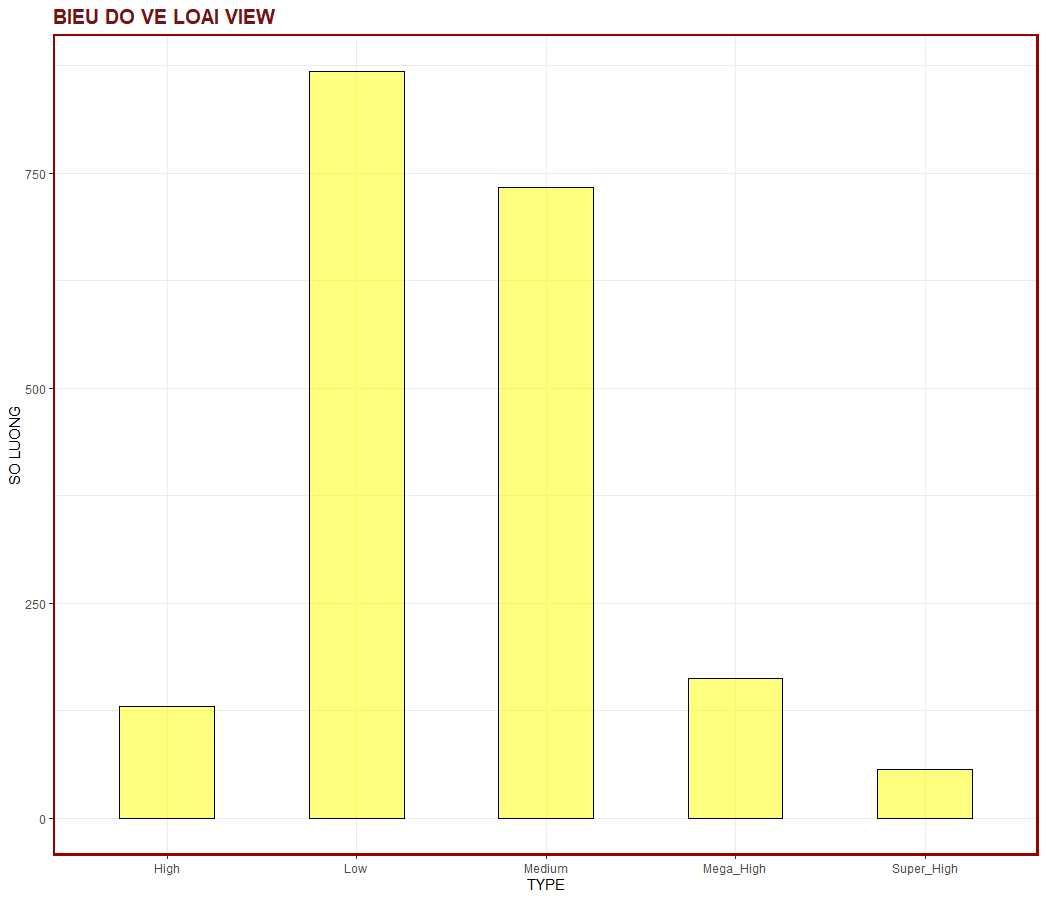
- Biểu đồ về tỉ lệ số từ

Về số từ

- Số truyện có lượng từ dưới 30.000 từ chiếm tỉ lệ cao nhất (55,38%)

- Số truyện có lượng từ trên 250.000 từ chiếm tỉ lệ thấp nhất (7.34%)

**- Plot 3:**

****

Nhận xét:

- Biểu đồ dạng cột

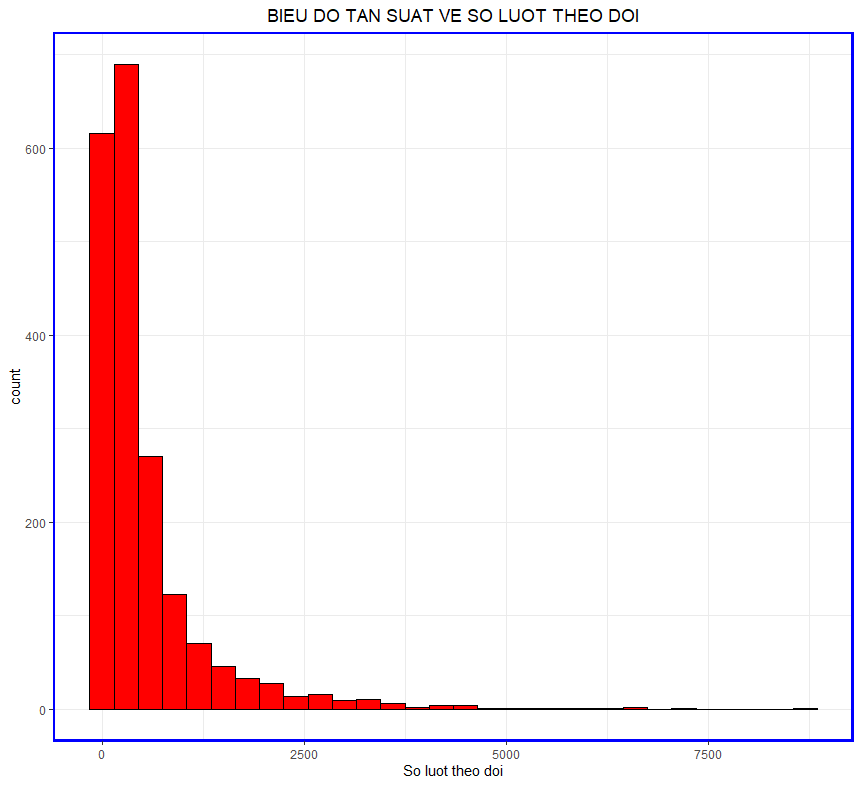
- Biểu đồ về số loại lượt xem

Về lượt xem:

- Số truyện có lượt xem dưới 30.000 lượt xem chiếm số lượng cao nhất: 868 truyện

- Số truyện có số lượt xem hơn 350.000 lượt xem chiếm số lượng thấp nhất: 57 truyện

**-Plot 4:**

****

Nhận xét:

- Biểu đồ histogram

- Biểu đồ về tỉ lệ số lượt theo dõi (có xu hướng giảm dần)

Về lượt theo dõi:

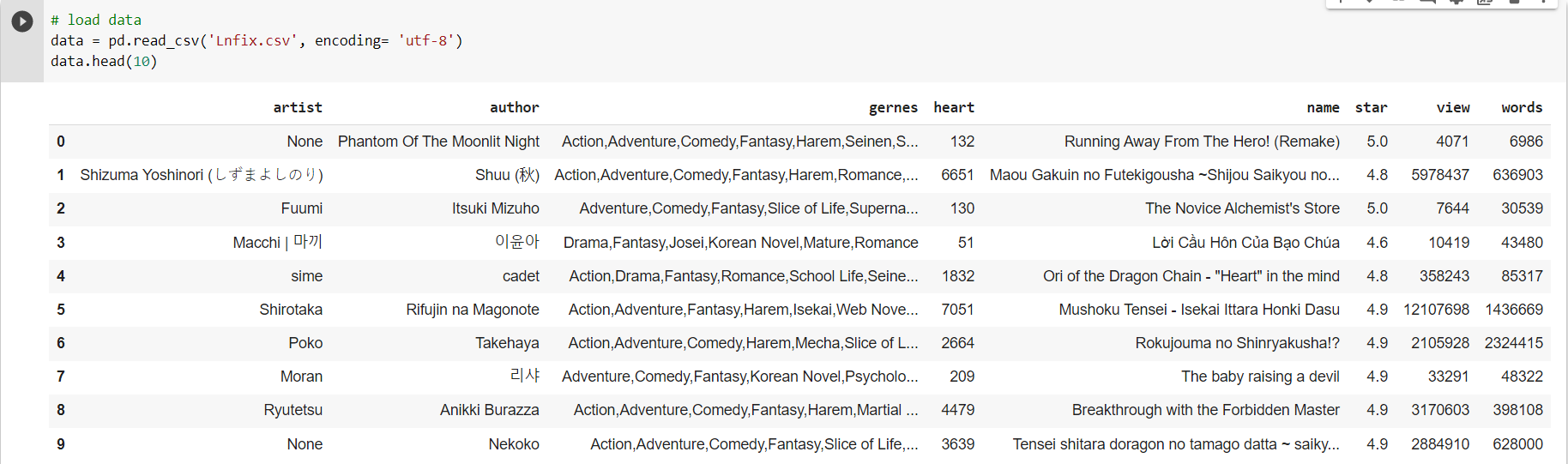
- Số truyện có lượt theo dõi từ khoảng 300 -> 600 chiếm số lượng cao nhất

- Số truyện có số lượt theo dõi trên 7000 chiểm số lượng ít nhất

# 6. Thống kê suy diễn

- Trong phần này, chúng ta sẽ sử dụng tệp dữ liệu “Lnfix.csv” để thực hiện thống kê suy diễn.

- Chúng ta cũng khai báo các thư viện và sử dụng hàm **read\_csv** để load dữ liệu vào Python.

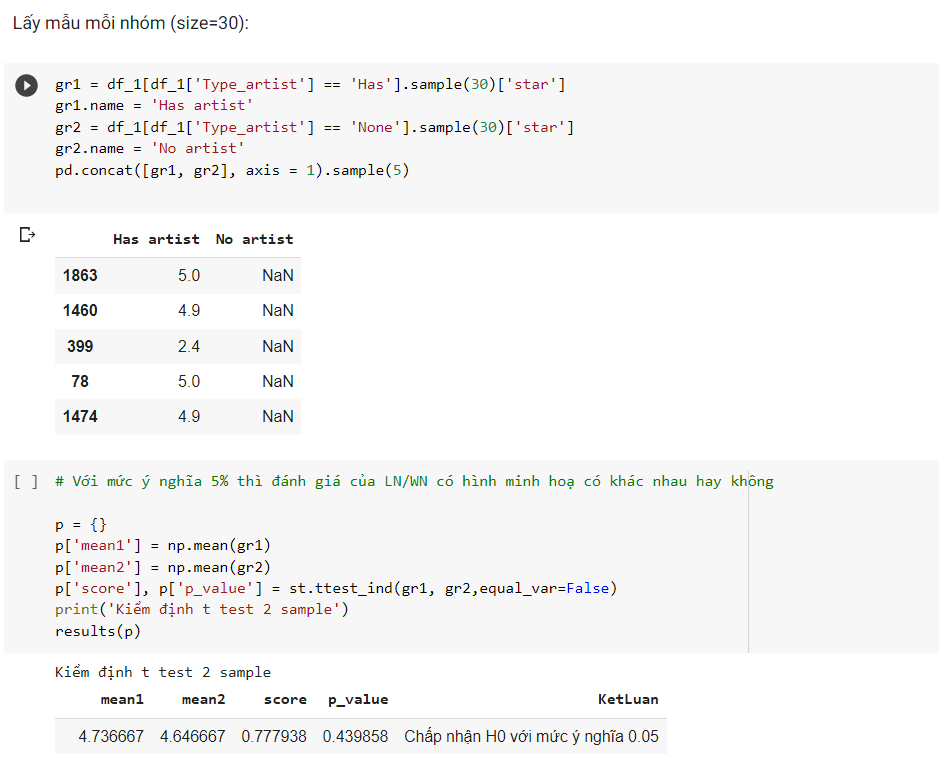


## A/ Kiểm định T-test

- Giả sử chúng ta cần kiểm định rằng LN/WN có hình minh hoạ và LN/WN không có hình minh hoạ có ảnh hưởng đến đánh giá hay không?

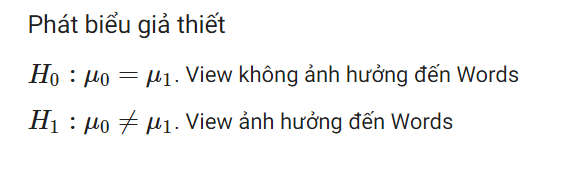
- Vậy chúng ta có giả thiết không và đối thiết như sau:

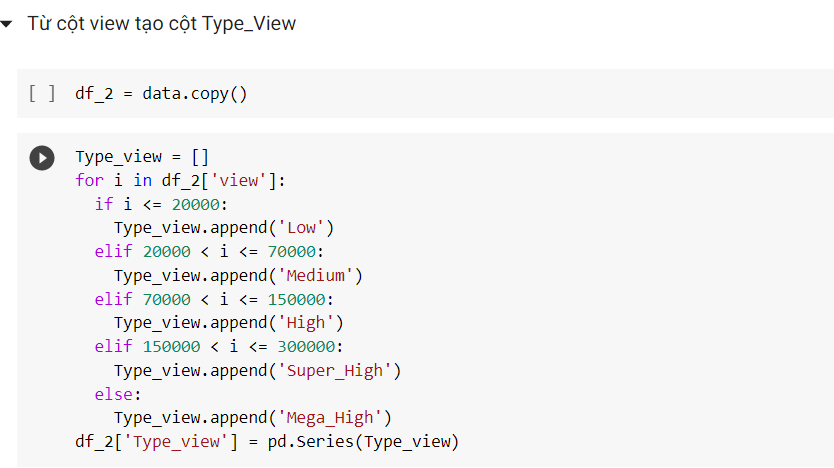
* H0: Đánh giá có minh hoạ hay không minh hoạ là như nhau
* H1: Đánh giá có minh hoạ hay không minh hoạ là khác nhau

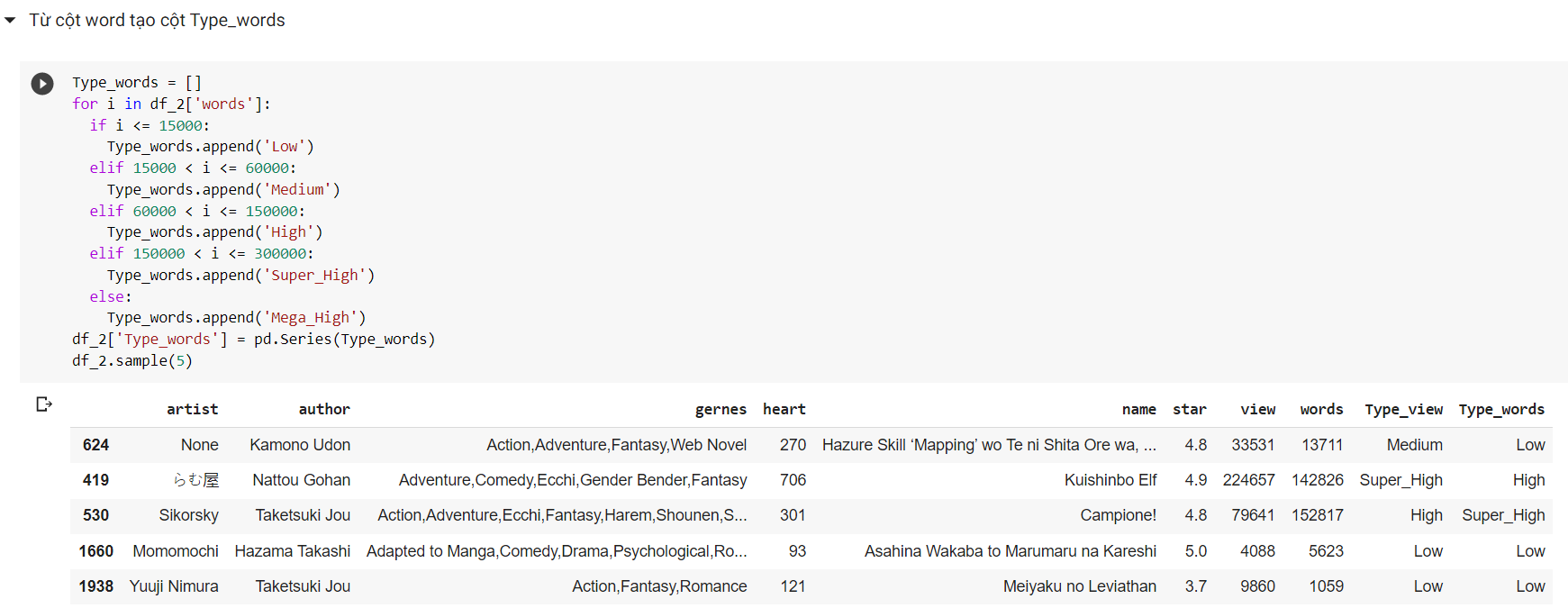


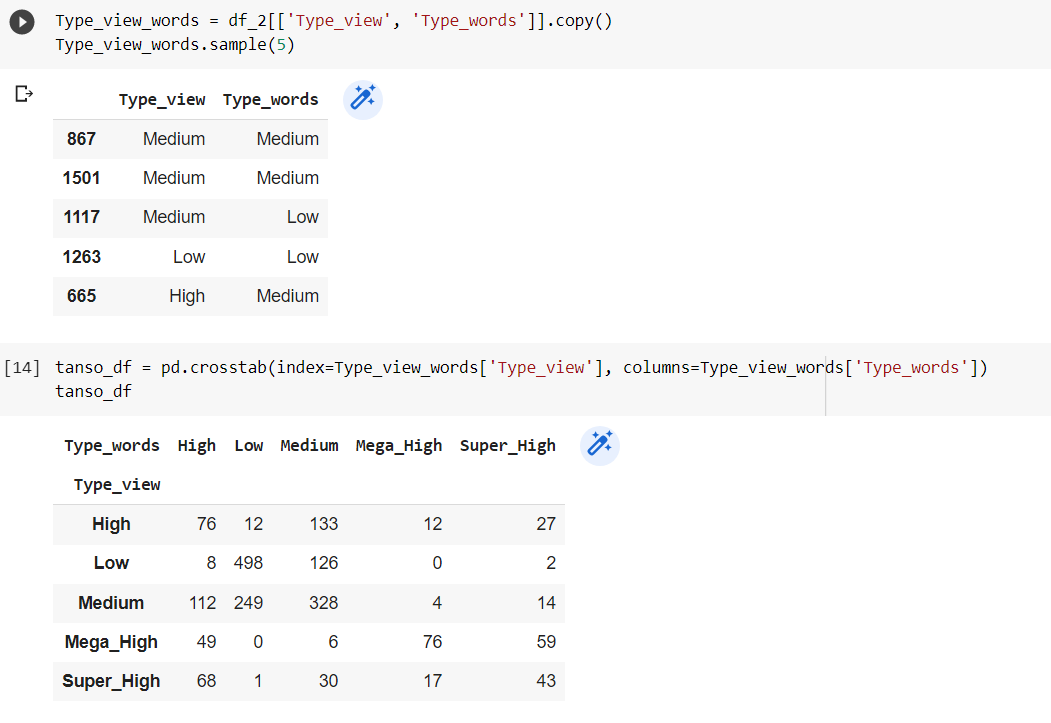
## B/ Kiểm định Chi2

### **B1. Dùng kiểm định chi2 để nghiên cứu ảnh hưởng của số lượt xem đến số chữ**





****

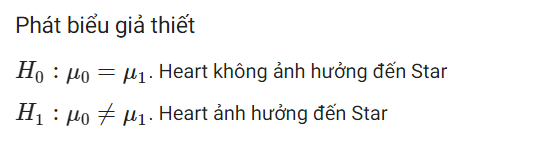
****

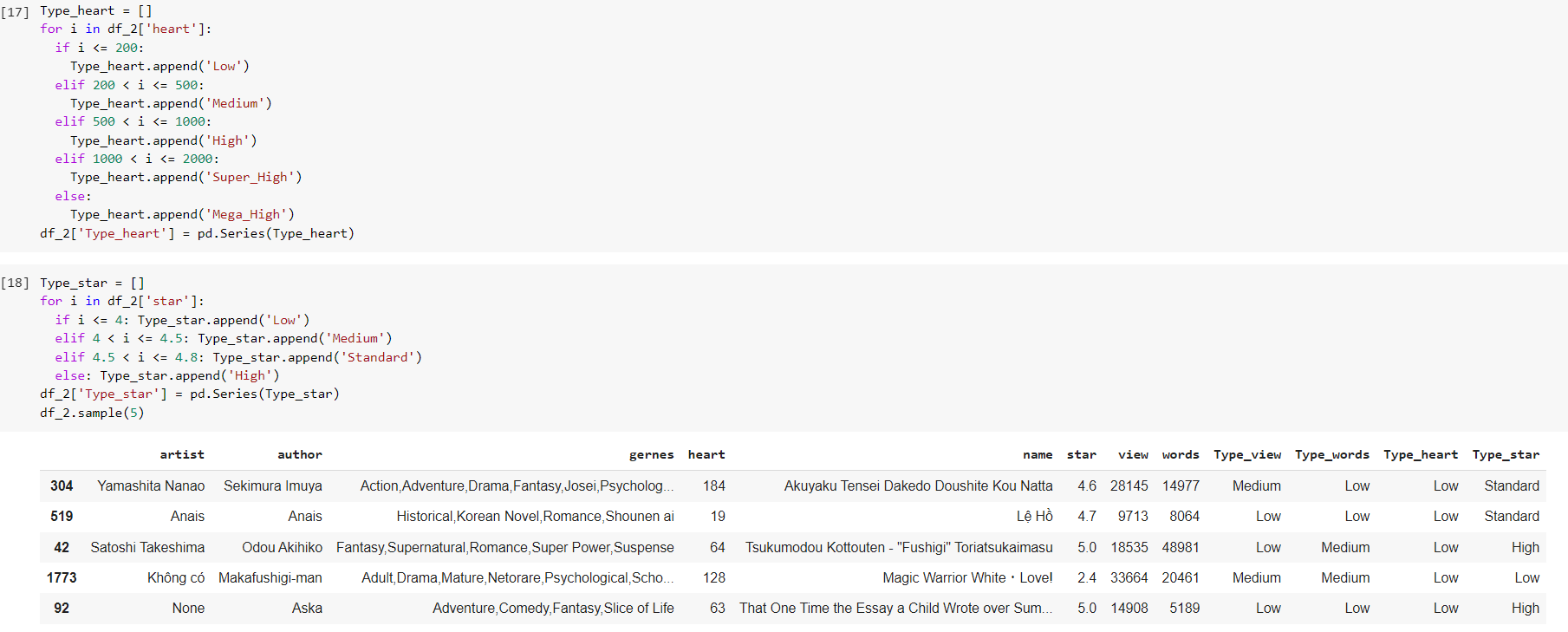
Sau các bước chỉnh sửa, phân tích dữ liệu của ‘view’ và ‘words’ ta tiến hành kiểm định Chi2

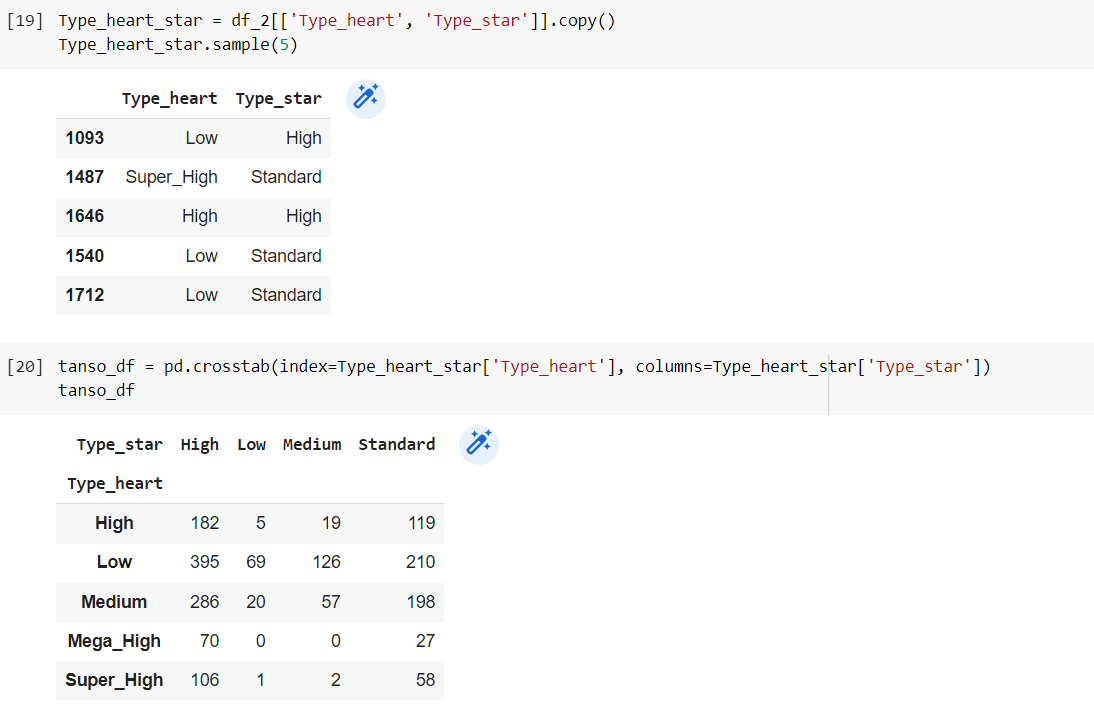


Với mức ý nghĩa 95% thì số lượt xem ảnh hưởng đến số chữ

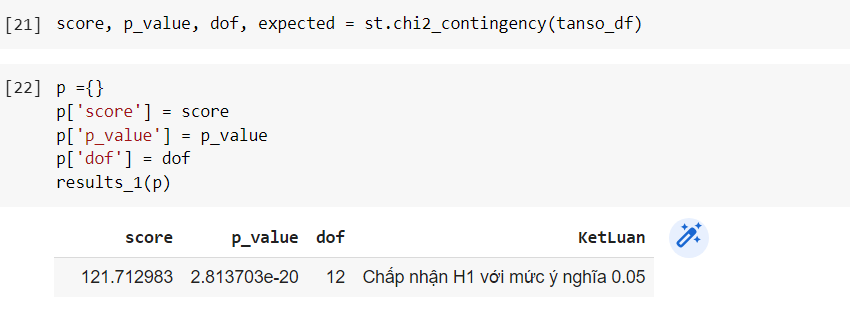
### **B.2/ Dùng kiểm định Chi2 để nghiên cứu ảnh hưởng của lượt theo dõi đến đánh giá**





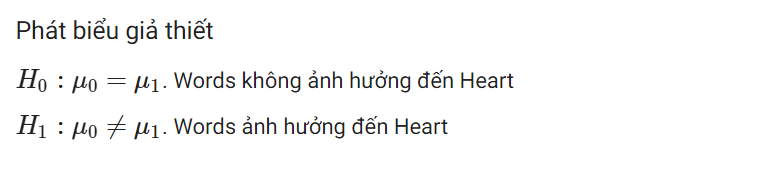


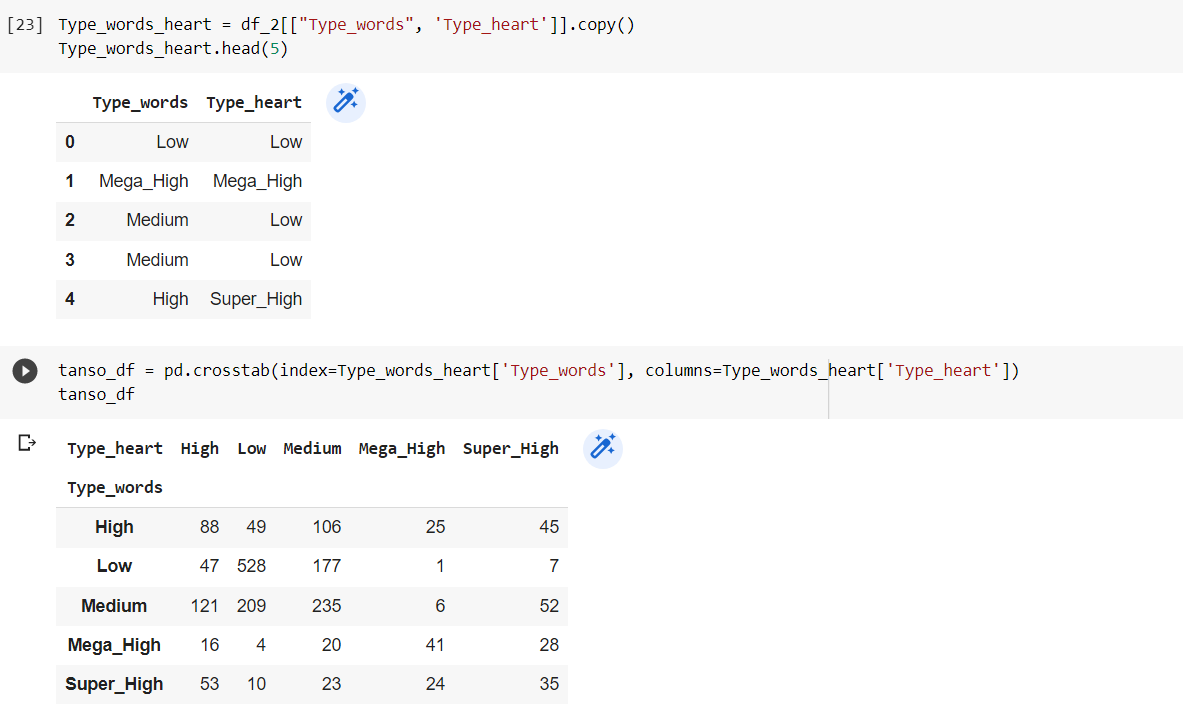
Sau các bước chỉnh sửa, phân tích dữ liệu của ‘heart’ và ‘star’ ta tiến hành kiểm định Chi2

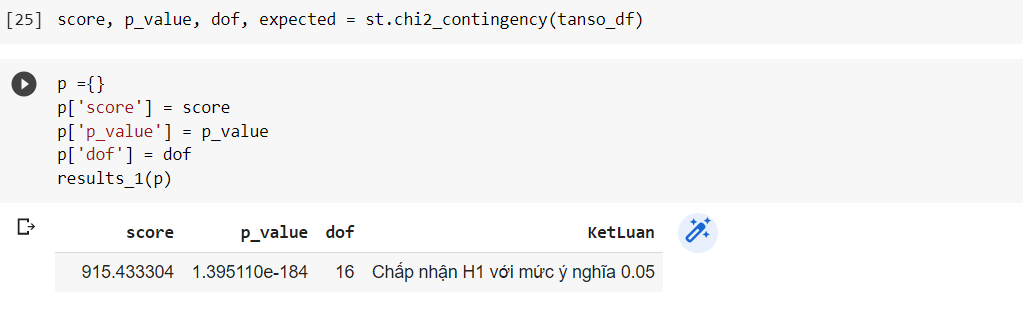


Với mức ý nghĩa 95% thì số lượt theo dõi ảnh hưởng đến đánh giá

### **B3/ Dùng kiểm định Chi2 để nghiên cứu ảnh hưởng của số chữ đến lượt theo dõi**







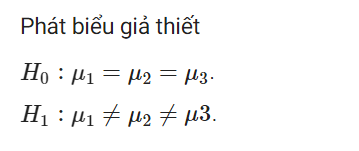
Với mức ý nghĩa 95% thì ‘heart’ ảnh hưởng đến ‘words’

## C/ Kiểm định kruskal-wallis

Ở đây ta tiến hành kiểm định ở 3 nhóm thể loại để xem có sự khác biệt về lượt theo dõi ở 3 nhóm thể loại này không.

3 nhóm thể loại là:

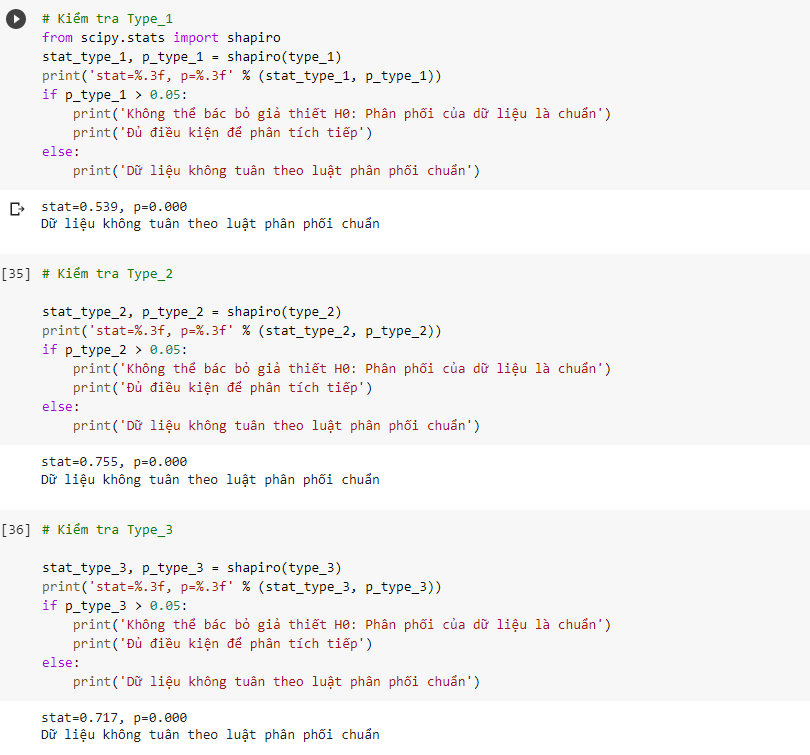
* Action – Comedy
* Action – Romance
* Comedy - Romance



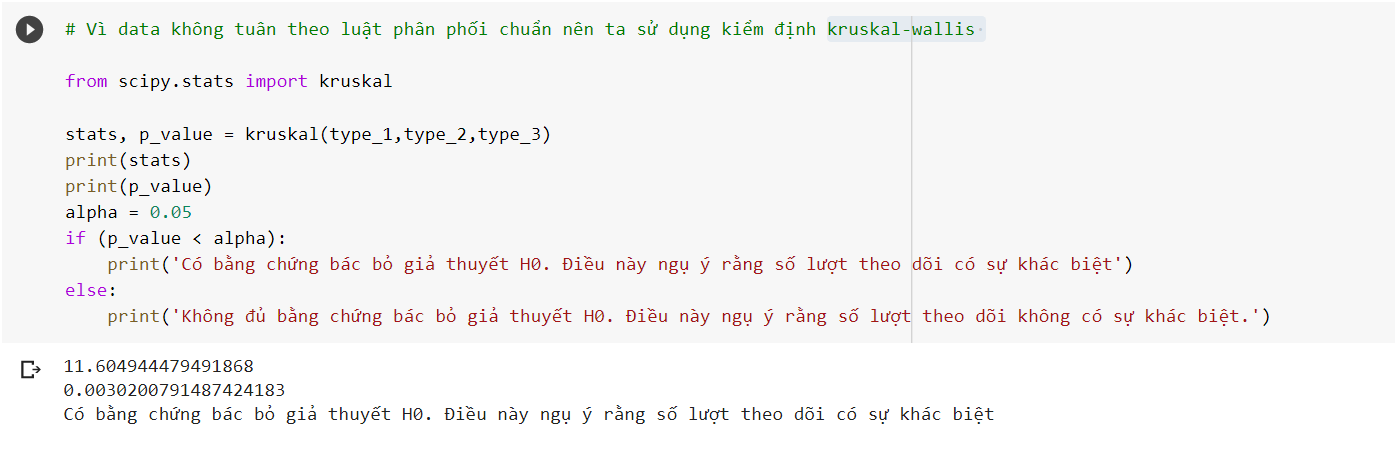


- Chú ý rằng, để thực hiện kiểm định ANOVA thì dữ liệu phải thỏa mãn các yêu cầu sau:

* Dữ liệu phải tuân theo phân phối chuẩn
* Dữ liệu đồng nhất về phương sai
* Các cá thể phải được lấy độc lập với nhau



Vì dữ liệu không chuẩn nên ta không thể dùng kiểm định phương sai Anova mà phải dùng kiểm định kiểm định phi tham số (kiểm định **kruskal-wallis)**

****

# 7. Tài liệu tham khảo

- Tham khảo cách sử dụng các thư viện và thống kê trong Python.

1. <https://pandas.pydata.org/docs/>
2. <https://matplotlib.org/>
3. <https://stackoverflow.com/>
4. <https://l.facebook.com/l.php?u=https%3A%2F%2Fphantichspss.com%2Fphan-tich-sau-anova-mot-yeu-post-hoc-one-way-anova.html%3Ffbclid%3DIwAR2lOChOYMWowu6o2_sFLUX-wFLoA7S9Lna82x68OiUFzT6Po9bATv5XKcc&h=AT17VQVFp9SC9u6NqnsD3NVNCd5vJtRspI5J7fJ6W1SrMkDHfnMGgDieernf-4sgILkaKZit32lUwotvtpbLsO3QK5XX4C6XcTAt9ooh2AdqorrjHBXBeDo2DpWC40g8DcV2-g>
5. <https://www.stat.uci.edu/what-is-statistics/>

<https://www.investopedia.com/terms/s/statistics.asp>