**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỀ TÀI THỰC TẬP CHUYÊN MÔN**

**ĐỀ TÀI**

**TÌM HIỂU VỀ R TRONG PHÂN TÍCH DỮ LIỆU**

Sinh viên thực hiện: Bùi Minh Đức

Mã Sinh Viên : 171202266

Lớp Công Nghệ Thông Tin 3 –K58

Khoa Công Nghệ Thông Tin

Người hướng dẫn: Th.s Bùi Ngọc Dũng

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỀ TÀI THỰC TẬP CHUYÊN MÔN**

**ĐỀ TÀI**

**TÌM HIỂU VỀ R TRONG PHÂN TÍCH DỮ LIỆU**

Sinh viên thực hiện :Bùi Minh Đức

Mã Sinh Viên : 171202266

Khoa Công Nghệ Thông Tin

Người hướng dẫn: Th.s Bùi Ngọc Dũng

Hà Nội , ngày 12 tháng 7 năm 2020

**Mục Lục**

[***Mở Đầu 4***](#_Toc44612026)

[***CHƯƠNG 1: Tổng quan về R 6***](#_Toc44612027)

***1.1 tổng quan R……………………………………………………………….……….6***

***1.1.2 R là gì …………………………………………………………………………..7***

***1.1.2.1 Lịch sử của R …………………………………………………………....……7***

***1.1.2.2 Tại sao nên học R …………………………………………………….………7***

***1.1.2.3 Đặc tình của R ………………………………………………………..…...…8***

***1.1.3 Ứng dụng của R* ……………………………………………………..………….*8***

[***CHƯƠNG 2: Cài đặt R 12***](#_Toc44612032)

[***CHƯƠNG 3: Thực hành với R 17***](#_Toc44612036)

[***CHƯƠNG* 4:*Cấu trúc dữ liệu R* 24**](#_Toc44612044)

[**Tài liệu tham khảo 33**](#_Toc44612055)

Lời cảm ơn

Để bài báo cáo hoàn thành một cách tốt đẹp , em xin gửi lời cảm ơn tới thầy giáo – Bùi Ngọc Dũng –đã quan tâm giúp đỡ , vạch hướng dẫn em hoàn thành một cách tốt nhất đồ án thực tập trong thời gian qua .Với điều kiện thời gian có hạn cũng như kinh nghiệm còn hạn chế của một sinh viên thực tập nên đồ án thực tập của em không tránh khỏi những thiếu xót . Em rất mong nhận được sự chỉ bảo , đóng góp ý kiến của thầy cùng toàn thể các bạn để nhóm chúng em có điều kiện bổ sung để em có thể hoàn thành đồ án thực tập hoàn chỉnh hơn .

Một lần nữa em xin chân thành cảm ơn !

Hà Nội , ngày 12 tháng 7 năm 2020

Sinh Viên Thực Tập

Bùi Minh Đức

Lời nói đầu

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin trong nhiều lĩnh vực của đời sống , kinh tế xã hội trong nhiều năm qua cũng đồng nghĩa với lượng dữ liệu đã được các cơ quan thu nhập và lưu trữ ngày một tích lũy nhiều lên . Họ lưu trữ các dữ liệu này vì cho rằng trong nó ẩn chưa những giá trị nhất định nào đó .Tuy nhiên theo thống kê chỉ có một lượng nhỏ (5%-10%) những dữ liệu này là luôn được phân tích , số còn lại họ không biết sẽ phải làm gì hoặc có thể làm gì với chũng nhưng họ vẫn tiếp tục thu thập rất tốn kém với ý nghĩ lo sợ rằng sẽ có cái gì đó quan trọng đã bị bỏ qua sau này có lúc cần đến nó , việc phân tích sử dụng dữ liệu tốt có ý nghĩa to lớn đến cá nhân, công ty và cả một đất nước .Vì vậy các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống không đấp ứng được thực tế từ đó đưa ra ngôn ngữ R giúp cho việc phân tích tổng hợp dữ liệu trở nên đơn giản hơn mà vẫn giữ nguyên được giá trị của dữ liệu .

**Hà Nội , ngày 12 tháng 7 năm 2020**

**Chương 1 : Giới thiệu chung về ngôn ngữ R**

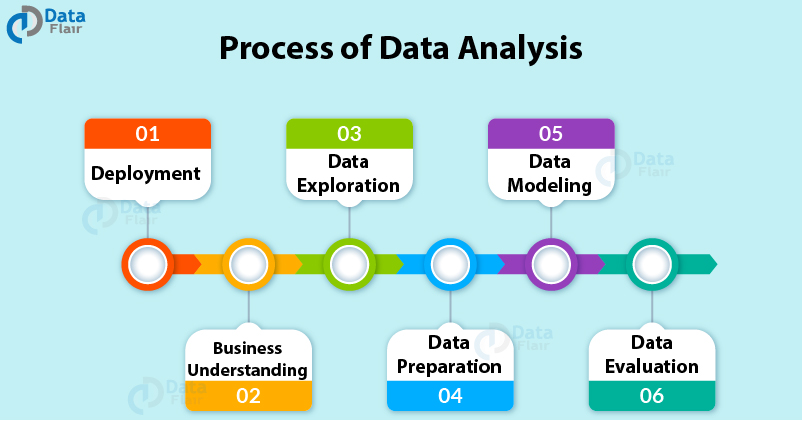
**1.1 . Tổng quan về R :**

R là một trong những ngôn ngữ được sử dụng rộng rãi cho mô hình thống kê .Nó được xem như là ngôn ngữ chung của Khoa học dữ liệu .Chúng ta thường nghe đến Data Analytics , Data Science – 2 xu hướng tốt nhất .Hiện nay , dữ liệu có vai trò quan trọng như dầu dùng trong ngành công nghiệp . Dữ liệu được thu thập ở dạng thô và được xử lý theo yêu cầu của một công ty và sau đó dữ liệu này được sử dụng cho mục đích ra quyết định. Quá trình này giúp các doanh nghiệp phát triển và mở rộng hoạt động trên thị trường. Nhưng, câu hỏi chính nảy sinh - quá trình được gọi là gì? Phân tích dữ liệu là câu trả lời ở đây. Nhà phân tích dữ liệu và Nhà khoa học dữ liệu là những người thực hiện quá trình này

**1.1.1 Data Analytics là gì ?**

Dữ liệu hoặc thông tin ở định dạng thô. Sự gia tăng kích thước của dữ liệu đã dẫn đến sự gia tăng nhu cầu thực hiện kiểm tra, làm sạch dữ liệu, chuyển đổi cũng như mô hình hóa dữ liệu để hiểu rõ hơn về dữ liệu để đưa ra kết luận cho quá trình ra quyết định tốt hơn. Quá trình này được gọi là phân tích dữ liệu

Quá trính thực hiện phân tích dữ liệu :



**1.1.1.1 Deployment** :

Chúng ta cần lên kế hoạch triển khai , giám sát và bảo trì và đưa ra báo cáo cuối cùng và đánh giá project . Trong giai đoạn này, chúng tôi triển khai các kết quả phân tích.Đây được xem là đánh giá dự án .Quy trình hoàn thành được xem như là quy trình phân tích kinh doanh .

**1.1.1.2 Business Understanding** :

Bất cứ khi nào có yêu cầu xảy ra, trước tiên chúng ta cần xác định mục tiêu kinh doanh, đánh giá tình hình, xác định mục tiêu khai thác dữ liệu và sau đó đưa ra kế hoạch dự án theo yêu cầu. Mục tiêu kinh doanh được xác định trong giai đoạn này

**1.1.1.3 Data Exploration** :

Đối với quá trình tiếp theo, chúng tôi cần thu thập dữ liệu ban đầu, mô tả và khám phá dữ liệu và cuối cùng xác minh chất lượng dữ liệu để đảm bảo dữ liệu đó chứa dữ liệu chúng tôi yêu cầu. Dữ liệu được thu thập từ các nguồn khác nhau được mô tả theo các ứng dụng của nó và sự cần thiết của dự án trong giai đoạn này. Điều này còn được gọi là thăm dò dữ liệu. Điều này là cần thiết để xác minh chất lượng của dữ liệu được thu thập

**1.1.1.1 Data Preparation** :

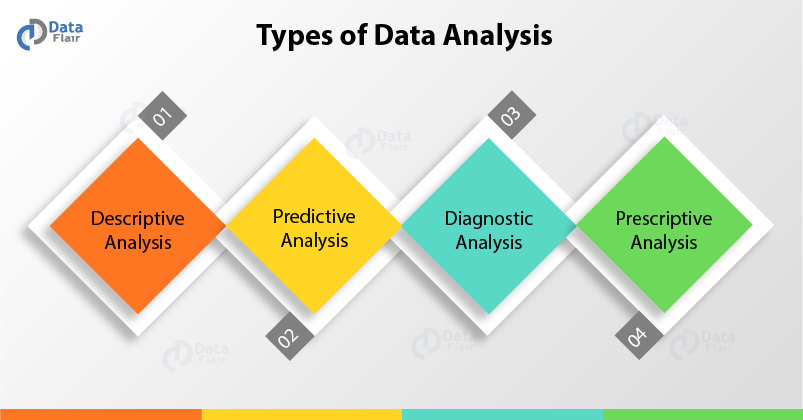
Từ dữ liệu được thu thập ở bước cuối cùng, chúng ta cần chọn dữ liệu theo nhu cầu, làm sạch nó, xây dựng nó để có được thông tin hữu ích và sau đó tích hợp tất cả. Cuối cùng, chúng ta cần định dạng dữ liệu để có được dữ liệu phù hợp. Dữ liệu được chọn, làm sạch và tích hợp vào định dạng được hoàn thiện để phân tích trong giai đoạn này.

**1.1.1.5 Data Modeling** :

Sau khi thu thập dữ liệu, chúng tôi thực hiện mô hình hóa dữ liệu trên đó. Đối với điều này, chúng ta cần chọn một kỹ thuật mô hình hóa, tạo thiết kế thử nghiệm, xây dựng mô hình và đánh giá mô hình được xây dựng. Mô hình dữ liệu được xây dựng để phân tích mối quan hệ giữa các đối tượng được chọn khác nhau trong dữ liệu. Các trường hợp thử nghiệm được xây dựng để đánh giá mô hình và mô hình được thử nghiệm và thực hiện trên dữ liệu trong giai đoạn này

**1.1.16 Data Evaluation** :

Ở đây, đánh giá kết quả từ bước cuối cùng, xem xét phạm vi lỗi và xác định các bước tiếp theo để thực hiện. Chúng tôi đánh giá kết quả của các trường hợp thử nghiệm và xem xét phạm vi lỗi trong giai đoạn này.



Các kĩ thuật sử dụng cho Data Analysis (Nguồn ảnh : data-flair.tranning)

**1.1.2 R là gì ?**

Là một ngôn ngữ lập trình mã nguồn mở tại điều kiên cho các thư viện đồ họa và tính toán thống kê. Là nguồn mở, R yêu thích sự hỗ trợ cộng đồng của các nhà phát triển khao khát làm việc phát hành các gói mới , cập nhật R và biến nó thành mã nguồn ổn định cho việc thực hành Khoa học dữ liệu .

Với sự giúp đỡ của R , người ta thực hiện các hoạt động thống kê khác nhau

Vì R là open-source nên hầu hết cấu trúc và thủ tục được phát triển bởi lập trình viên trên toàn thế giới . Tất cả các gói đều có sẵn và free trên website R được gọi là CRAN.

Nó gồm hơn 10000 gói .Tiến trình cài đặt cơ bản được dựng sẵn , nhiều dữ liệu khoa học và thống kê đế sử dụng cho nhiều tác vụ .

**1.1.2.1 Lịch sử của R :**

R được định nghĩa ở Bell Laboratories bởi John Chambers năm 1976.

R được phát triển và mở rộng dựa trên ngôn ngữ S .

Dự án R được phát triển bởi Ross Ihaka và Robert Gent quý ông và phát hành năm 1992, phiên bản đầu tiên vào năm 1995 và phiên bản beta ổn định vào năm 2000.

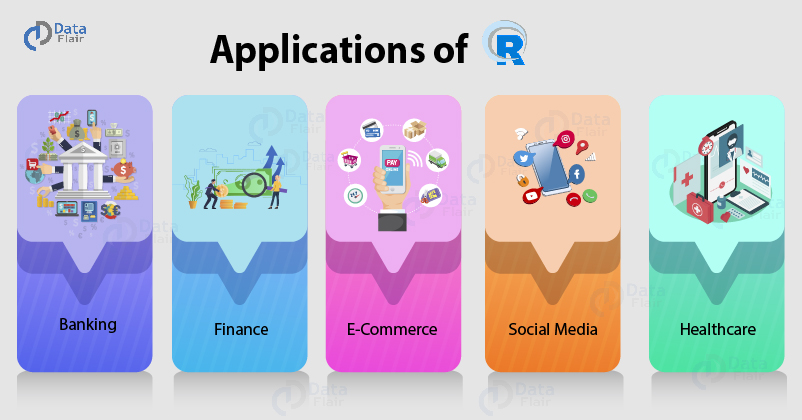
**1.1.2.2 Tại sao nên học R**

* Với R, bạn có thể thực hiện phân tích thống kê, phân tích dữ liệu cũng như học máy. Chúng ta có thể tạo các đối tượng, hàm và gói trong đó. R độc lập với nền tảng và có thể được sử dụng trên nhiều hệ điều hành. R miễn phí nhờ cấp phép GNU nguồn mở và có thể được cài đặt bởi bất kỳ ai
* R bao gồm một bộ sưu tập mạnh mẽ của các thư viện đồ họa như ggplot2, cốt truyện và nhiều hơn nữa. Với các thư viện này, bạn có thể thực hiện trực quan hấp dẫn và trực quan thanh lịch.
* R được sử dụng rộng rãi nhất bởi các ngành công nghiệp khác nhau. Chỉ có các con đường học thuật trong quá khứ đã sử dụng R nhưng các ngành công nghiệp hiện đang sử dụng R làm công cụ chính cho mô hình thống kê. Ngành công nghiệp sâu sắc nhất sử dụng R là ngành Khoa học dữ liệu và một số ngành công nghiệp cơ bản mà nó bao gồm. các ngành công nghiệp như y tế, tài chính, ngân hàng, sản xuất và nhiều hơn nữa
* Có khoảng 2 triệu cơ hội việc làm cho các lập trình viên R trên toàn thế giới. Các công ty thuê các lập trình viên R cho nhiều vai trò như nhà phân tích dữ liệu, nhà phân tích kinh doanh, chuyên gia trực quan hóa dữ liệu và chuyên gia tình báo kinh doanh

**1.1.2.3 Đặc tình của R :**

* R là một ngôn ngữ lập trình toàn diện, cung cấp hỗ trợ cho lập trình thủ tục liên quan đến các chức năng cũng như lập trình hướng đối tượng với các chức năng chung.
* Có hơn 10.000 gói trong kho lưu trữ lập trình R. Với các gói này, người ta có thể sử dụng các chức năng để tạo điều kiện cho việc lập trình dễ dàng hơn.
* Là một ngôn ngữ dựa trên trình thông dịch, R tạo ra một mã độc lập với máy có thể mang theo được trong nhiều môi trường. Hơn nữa, nó tạo điều kiện dễ dàng gỡ lỗi các lỗi trong mã
* R tạo điều kiện cho các hoạt động phức tạp với các vectơ, mảng, khung dữ liệu cũng như các đối tượng dữ liệu khác có kích thước khác nhau.
* R có thể dễ dàng tích hợp với nhiều công nghệ và khung khác như Hadoop và HDFS. Nó cũng có thể tích hợp với các ngôn ngữ lập trình khác như C, C ++, Python, Java, FORTRAN và JavaScript
* R cung cấp các phương tiện mạnh mẽ để xử lý và lưu trữ dữ liệu.
* Như đã thảo luận trong phần trên, R có hỗ trợ cộng đồng rộng lớn cung cấp hỗ trợ kỹ thuật, hội thảo và một số trại khởi động để giúp bạn bắt đầu với R.
* R tương thích đa nền tảng. Các gói R có thể được cài đặt và sử dụng trên mọi HĐH trong mọi môi trường phần mềm mà không có bất kỳ thay đổi nào

**1.1.3 Ứng dụng của R :**



* **Academics: R là một ngôn ngữ lập trình được tạo ra cho tính toán thống kê. Nó được sử dụng rộng rãi bởi các sinh viên, giáo sư và nhà nghiên cứu để phát triển các mô hình thống kê và cũng để phân tích kết quả lớn từ các nghiên cứu và khảo sát**
* **Healthcare :** **R là công cụ phân tích phổ biến nhất được sử dụng cho các thử nghiệm thuốc tiền lâm sàng. Các loại thuốc và quy trình y tế mới được thử nghiệm bằng R. Nó cũng được sử dụng để phân tích trình tự di truyền để xác định dị thường di truyền và phân tích các thành phần hóa học khác nhau để xác định công dụng và tác dụng phụ của chúng**
* **Finance :** **Lĩnh vực tài chính liên quan đến rất nhiều số liệu thống kê. R là công nghệ được ưa thích nhất cho mục đích này trong số đó. Họ sử dụng nó để xác định và dự đoán xu hướng thị trường. Họ cũng sử dụng nó để tính toán các số liệu thống kê như lãi hoặc lỗ hàng năm, mô hình phục hồi, v.v**
* **Banking :** **Ngân hàng sử dụng R để mô hình hóa rủi ro và tính toán lãi suất thế chấp. Họ cũng sử dụng nó cho mô hình rủi ro tín dụng**
* **IT Sector :** **Các công ty CNTT sử dụng R cho trí tuệ kinh doanh và để phát triển phần mềm tính toán thống kê. Họ cũng sử dụng R để đưa ra chiến lược kinh doanh**
* **Manufacturing :** **Các công ty sản xuất sử dụng R để dự đoán nhu cầu thị trường. Họ thay đổi sản xuất của họ theo dự đoán nhu cầu. Họ cũng thiết kế sản phẩm của họ theo kết quả phân tích tình cảm đánh giá của khách hàng**
* **Social Media :** **Các công ty truyền thông xã hội sử dụng R cho nghiên cứu máy học và phân tích tình cảm. Họ phân tích các bài đăng và hoạt động của người dùng của họ để rút ra những hiểu biết sâu sắc từ nó và cố gắng cải thiện trải nghiệm người dùng của họ theo kết quả**

**Chương 2 : Cài đặt R & RStudio**

**2.1.Cài đặt R :**

**Cài đặt cho Windows**

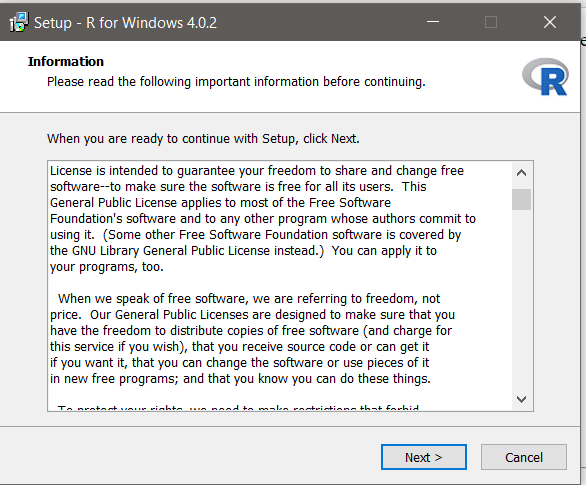
Bước 1 : Truy cập Website ***https://cran.rproject.org/bin/windows/base/***

Bước 2 : Chọn Download R 4.0.2 for Windows

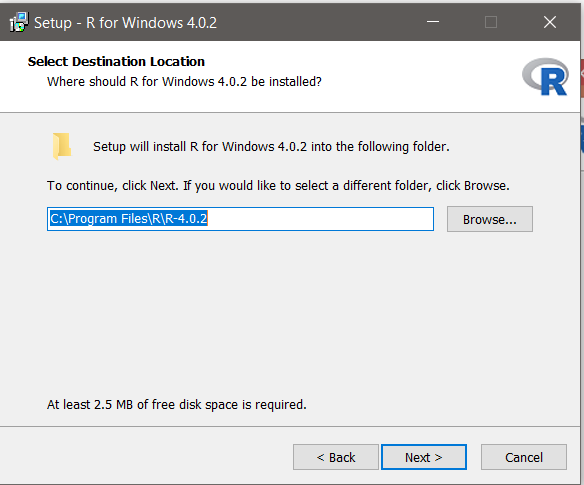


Bước 3 : Click chuột vào thư mục vừa tải về

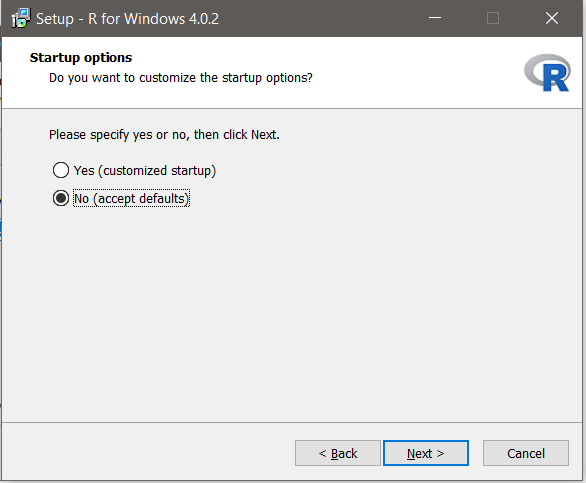
Bước 4 : Chọn Next



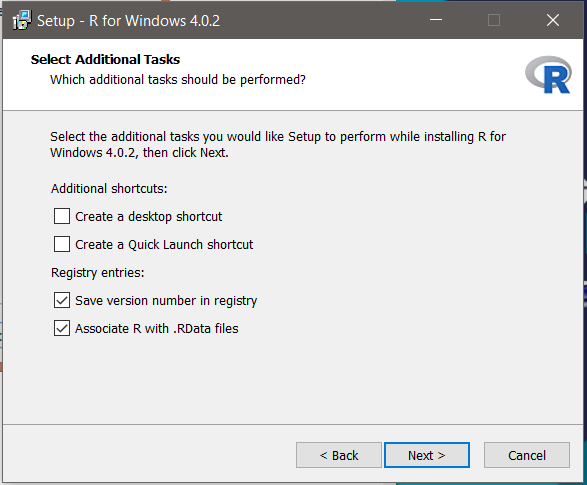
Bước 5: Chọn thư mục cần lưu trữ



Bước 6 : Trong bước tiếp theo, bạn sẽ được hỏi nếu bạn muốn tùy chỉnh khởi động. Nhấp vào Không (chấp nhận mặc định)



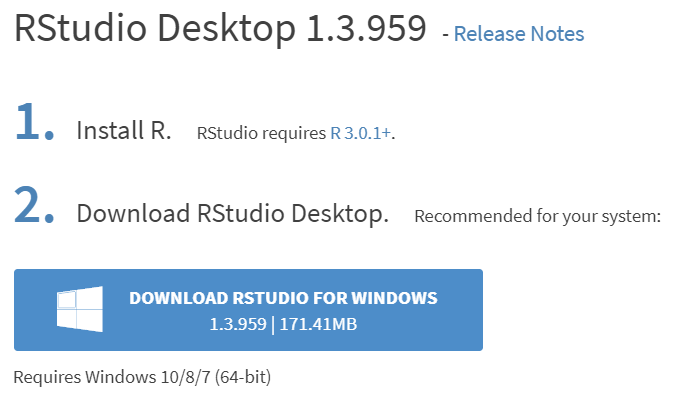
Bước 7 : Trong bước tiếp theo, bạn sẽ được yêu cầu thêm chi tiết. Bạn có thể tạo lối tắt trên màn hình nền hoặc Khởi động nhanh bằng cách chọn các hộp



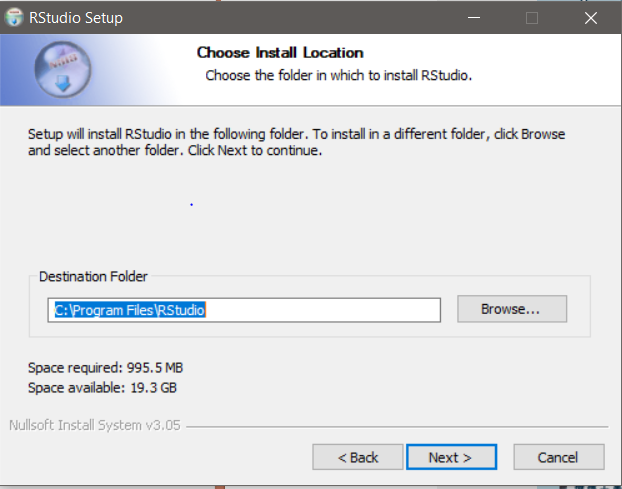
**2.2 Cài đặt Rstudio :**

Bước 1 : Truy cập vào <https://www.rstudio.com/products/rstudio/download/>

Bước 2: Ấn chọn DownLoad Rstudio :



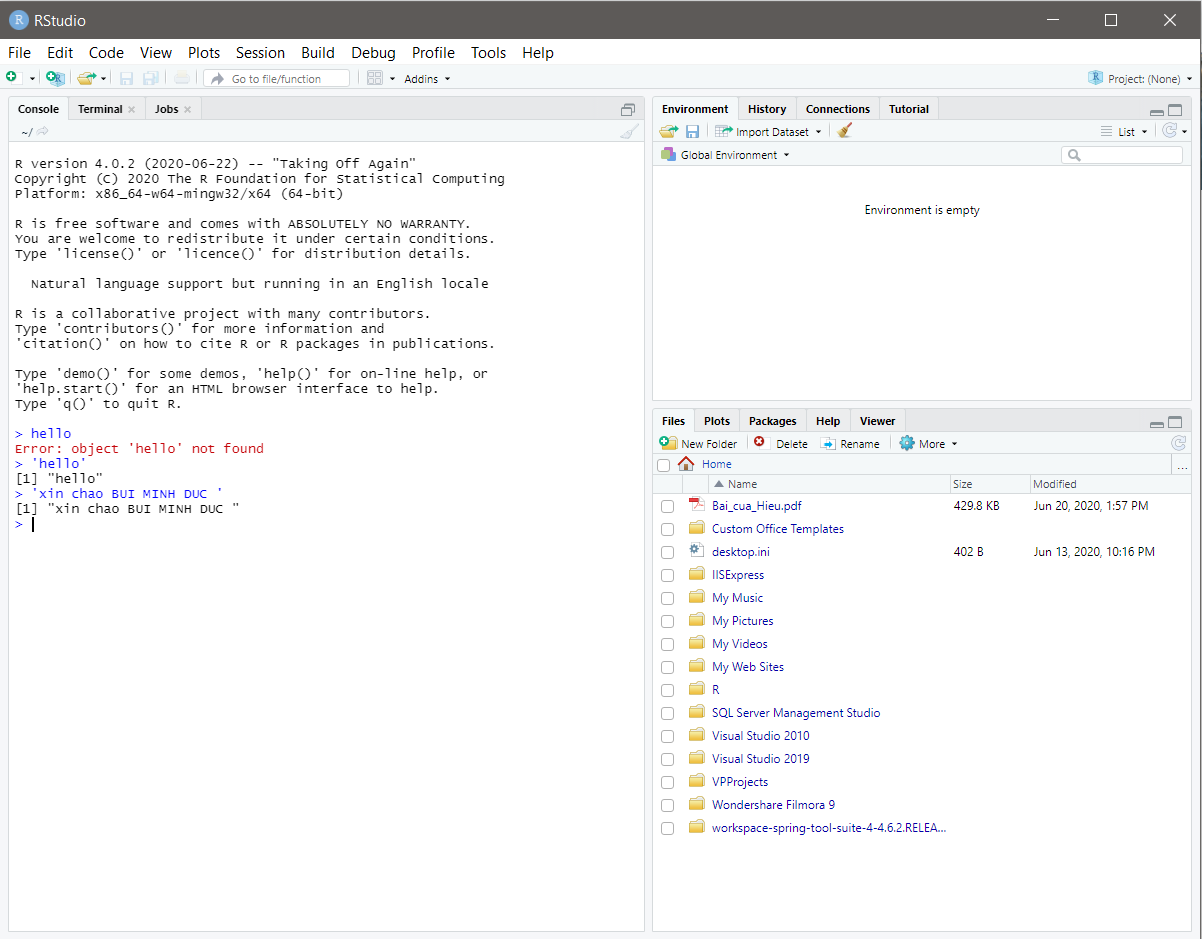
Bước 3 : Cài đặt



Bươc 4 : Quá trình cài đặt đang thực thi :



Bước 5 : Cài đặt xong và mở ứng dụng :

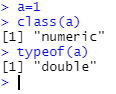


**Chương 3 : Thực hành với R**

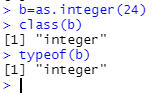
**3.1 Kiểu dữ liệu :**

**3.1.1 Kiểu dữ liệu trong R là gì ?**

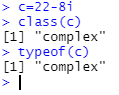
* Để xử lý chính xác, một ngôn ngữ lập trình phải biết những gì có thể và không thể được thực hiện đối với một giá trị cụ thể. Ví dụ, bổ sung không thể được thực hiện trên các từ ‘xin chào, và world thế giới.
* Tương tự, bạn không thể thay đổi số 1 và -34,5 từ chữ thường sang chữ hoa. Do đó, R có một tính năng gọi là kiểu dữ liệu. Các loại giá trị khác nhau được gán các loại dữ liệu khác nhau giúp phân biệt chúng. Các loại này có các đặc điểm và quy tắc nhất định liên quan đến chúng xác định các thuộc tính của chúng.
* R cung cấp các hàm class () và typeof () để tìm ra lớp và loại của bất kỳ biến nào. R có năm loại dữ liệu là:
* Numeric



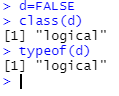
* Integers



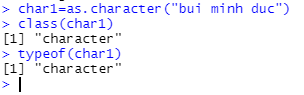
* Complex



* Logical

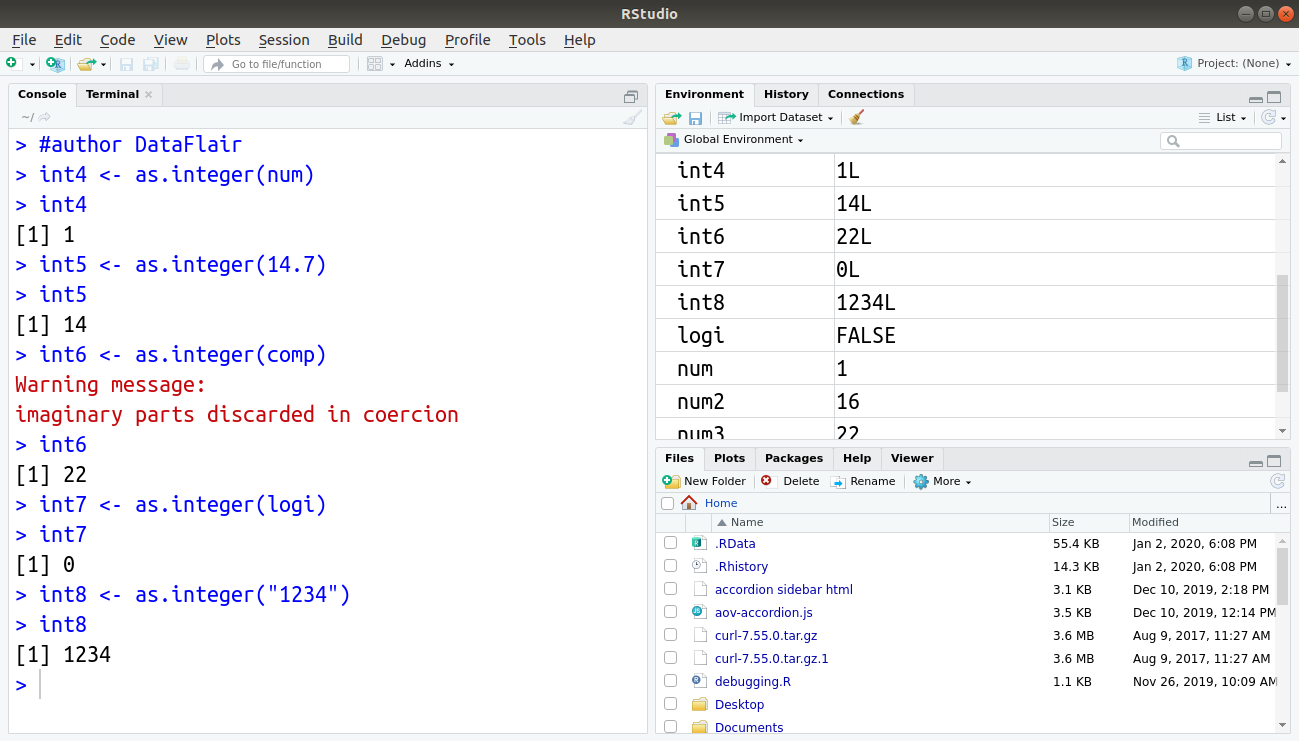


* Characters



**3.1.2, Đổi kiểu trong R :**

**Sử dụng [ as.’Tên kiểu cần đổi ’ ]**



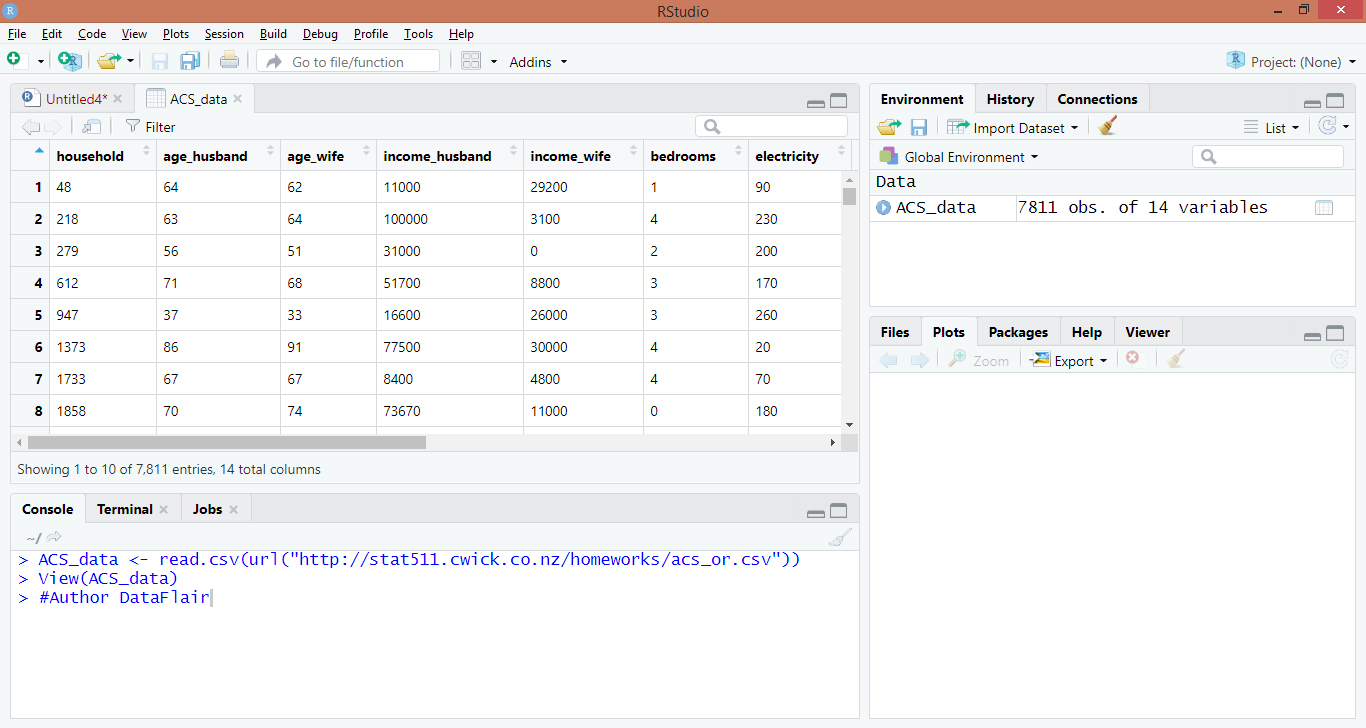
**3.2 .Import dữ liệu trong Rstudio** :

Để đọc dữ liệu sử dụng cú pháp : read.csv(“file.csv”)

Ví dụ :

read.**csv**(**url**("http://stat511.cwick.co.nz/homeworks/acs\_or.csv"))

View() : Hiển thị toàn bộ dữ liệu của bảng

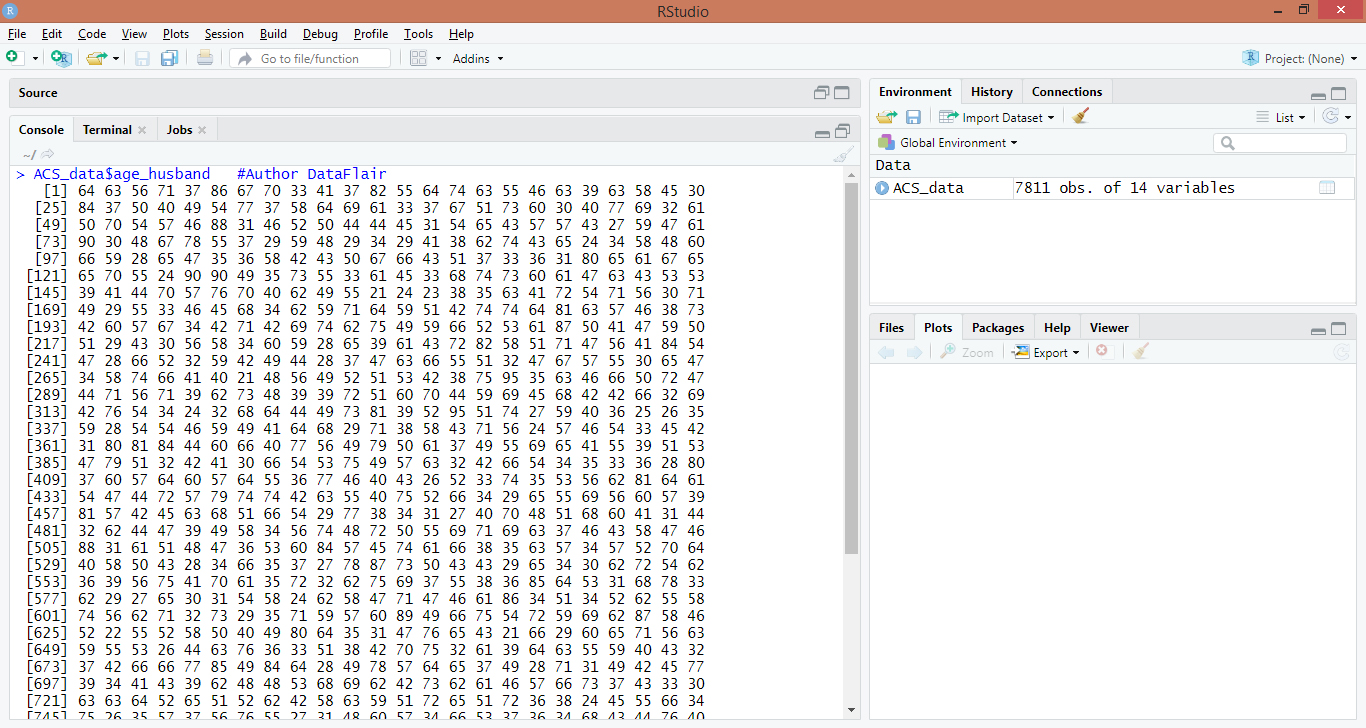


**3.2.1** , Cách lấy dữ liệu theo người dùng :

**3.2.1.1** Lấy tất cả giá trị của cột age\_husband

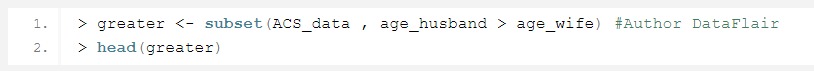
Cú pháp :

***Tên\_Bảng$Tên\_Cột***

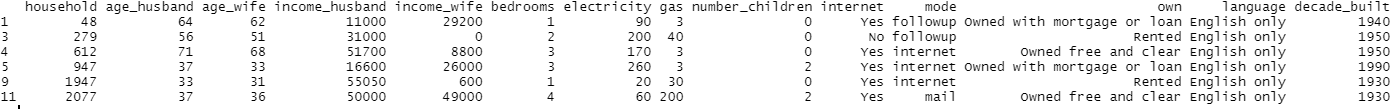


**3.2.1.2**

\* Nếu bạn muốn truy xuất dữ liệu mà age\_husband lớn hơn nhãn age\_wife, thì chúng tôi sẽ thực hiện lệnh sau :

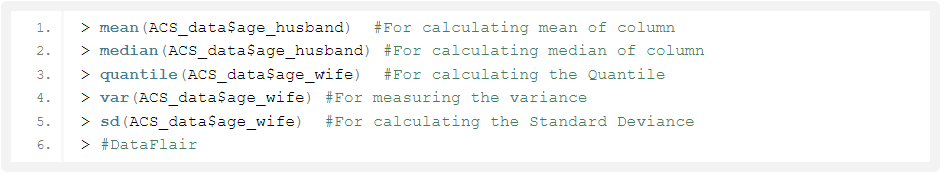


Kết quả :



* + Hàm subset tỏ ra hiệu quả trong trường hợp này

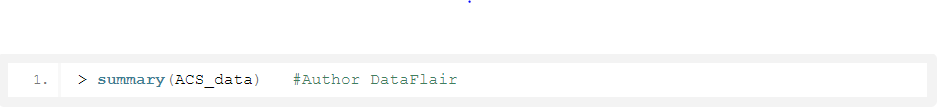
**3.2.1.3 Một số phương thức khác :**

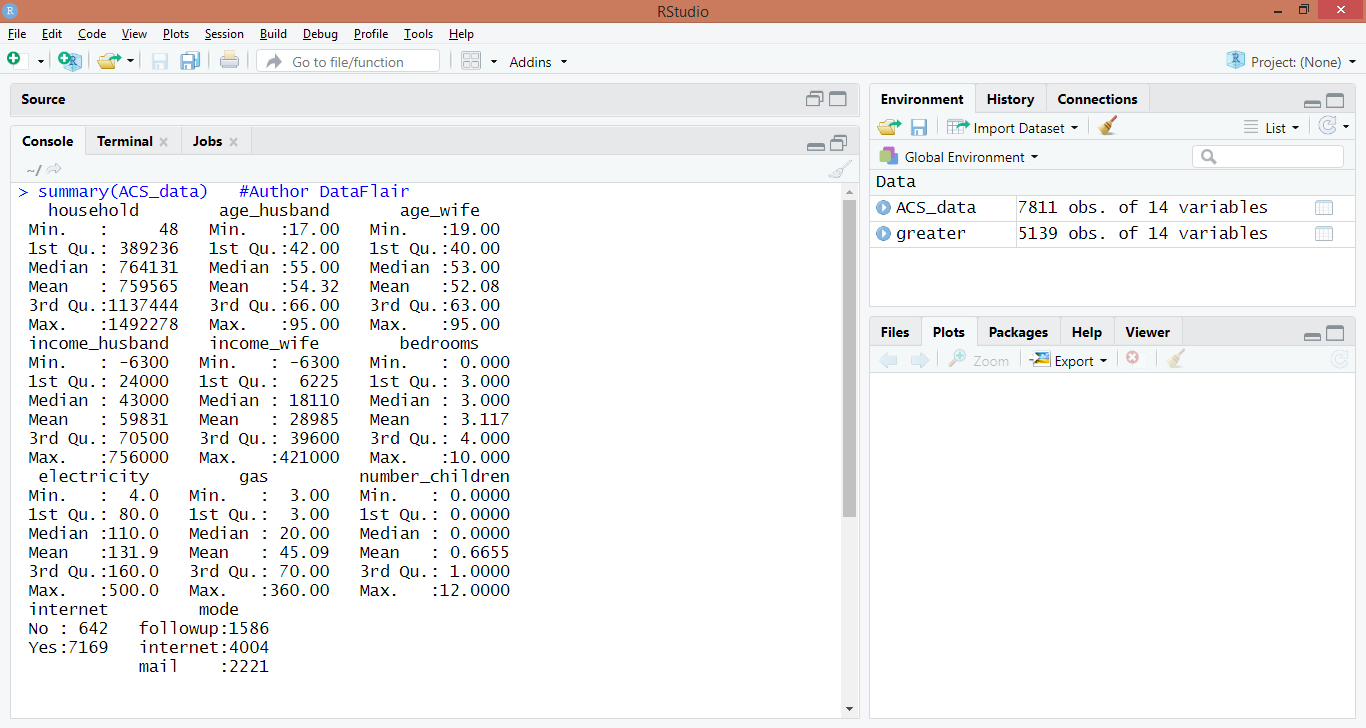


* Median : tính giá trị trung bình
* Quantitle:định lượng
* Var : Tính phương sai
* Sd: tính độ lệch chuẩn



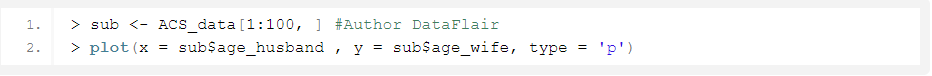
**3.2.1.4 Lấy bản tóm tắt của bộ dữ liệu :**



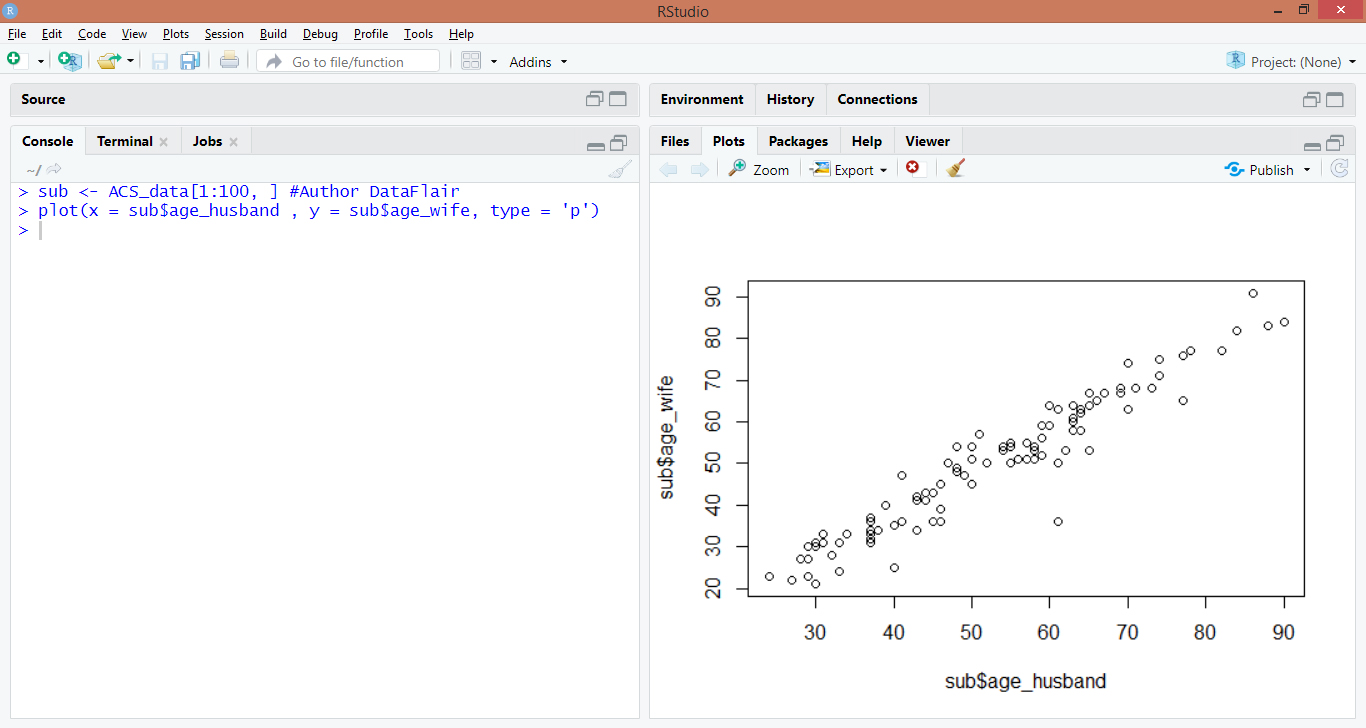


**3.3 Plotting Data by Rstudio :**

* *RStudio cung cấp các tính năng trực quan hóa đồ họa tiên tiến. Chúng ta có thể vẽ dữ liệu trên bằng nhãn cột ‘age\_husband, trên trục x và cột age\_wife trên trục y .*
* Vẽ đồ thị mỗi tương quan giữa trường age\_husband và age\_wife

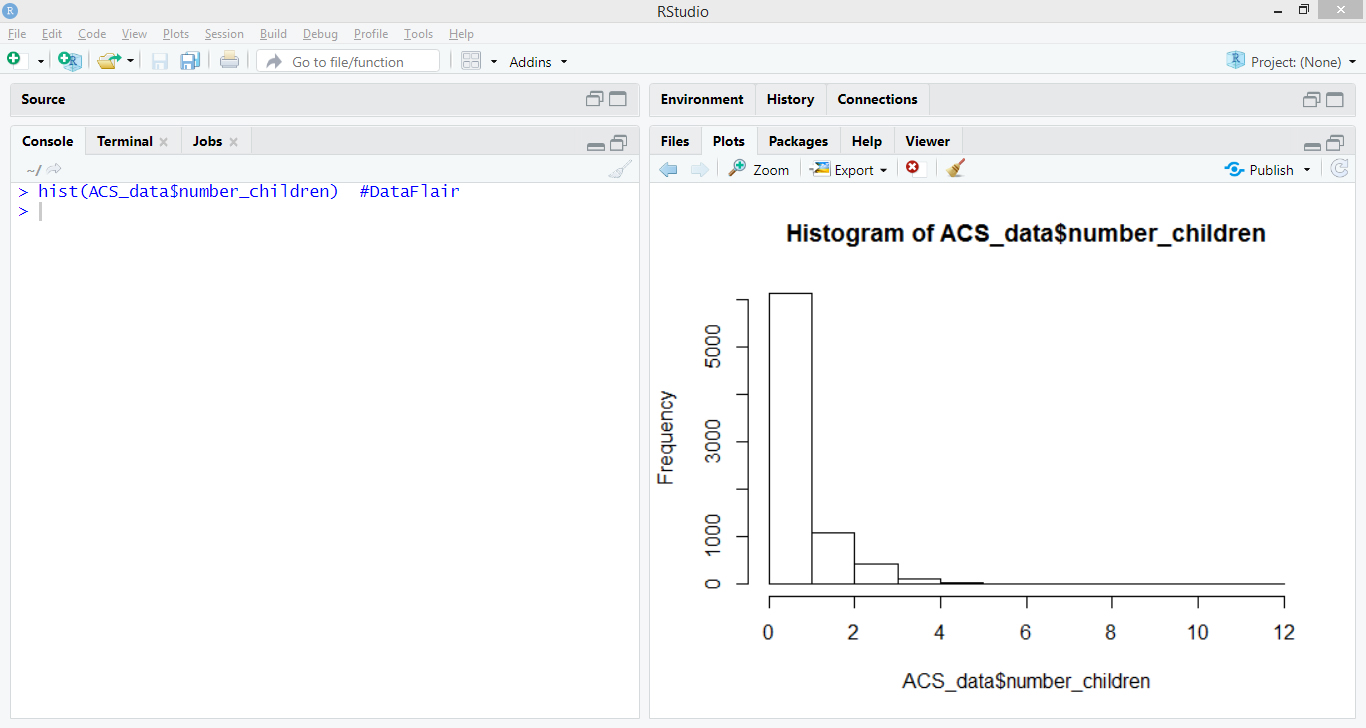


**Output :**



* Vẽ biểu đồ cột cho trường $number\_children





**Từ đây có thể thấy rằng số lượng gia đình có 1-2 con chiếm số lượng lớn**

**Chương 4 Cấu trúc dữ liệu R**

* 1. **Vector :**

**4.1.1 Vector là gì ?**

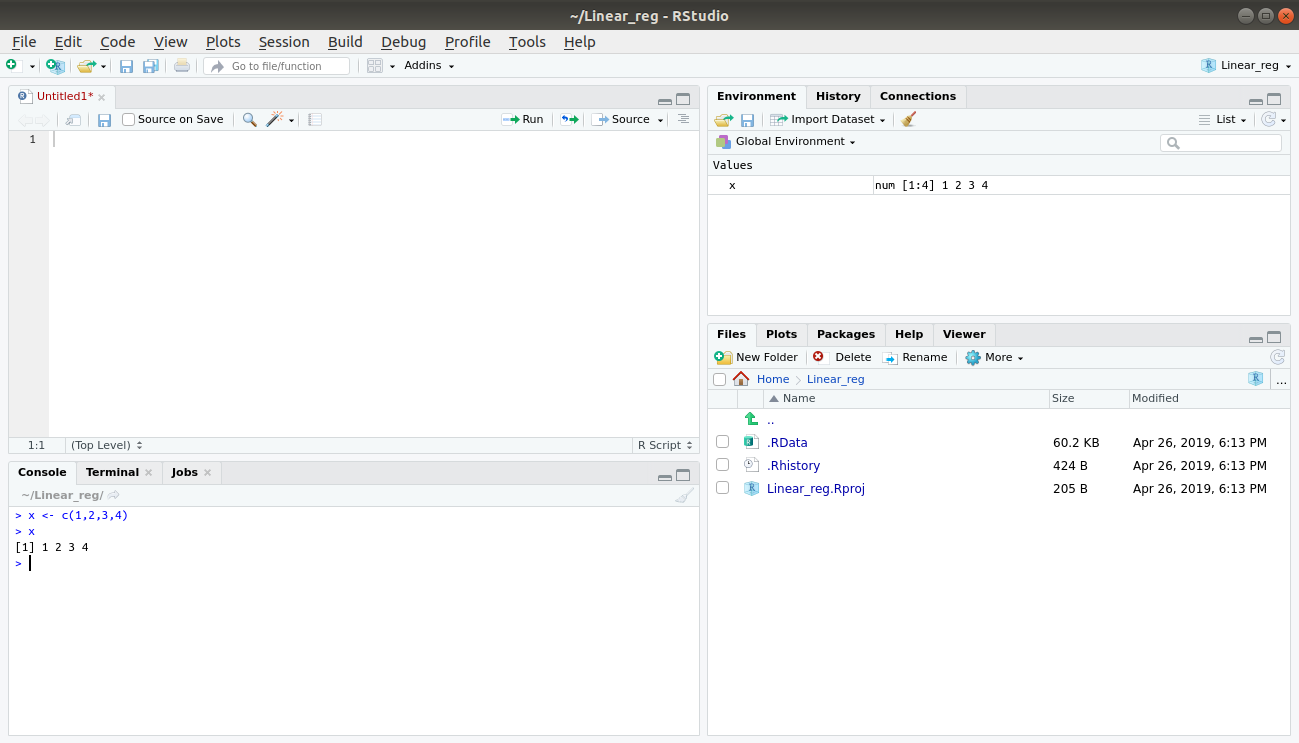
* Là chuỗi các phần tử có chung kiểu dữ liệu , những nguyên tố này gọi là thành phần của vector
* R Vector gồm 2 phần : ***Atomic vectors*** và ***Lists*** .Chúng có 3 đặc tính chung :
* Kiểu của hàm
* Độ lớn
* Phương thức thuộc tính

Chúng có điểm khác biệt : kiểu của phần tử , các elements trong vector phải cùng kiểu còn trong Lists thì không

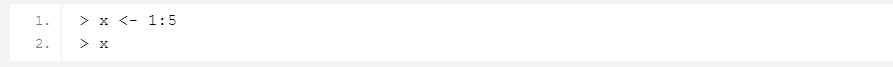
**4.1.2 Cách tạo Vector**

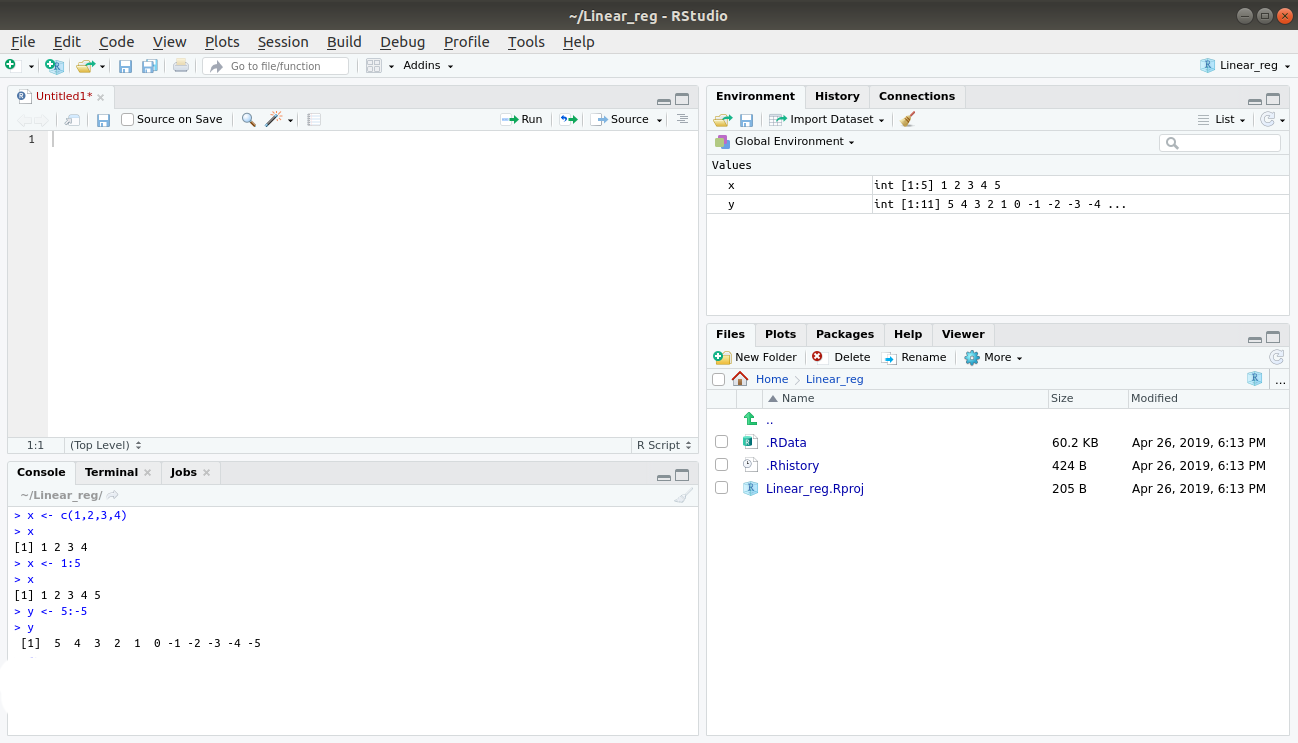
Hàm *c()* để tạo vector trong R , nó trả về mảng 1 chiều





* + 1. **Sử dụng toán tử**

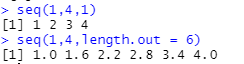


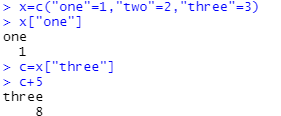


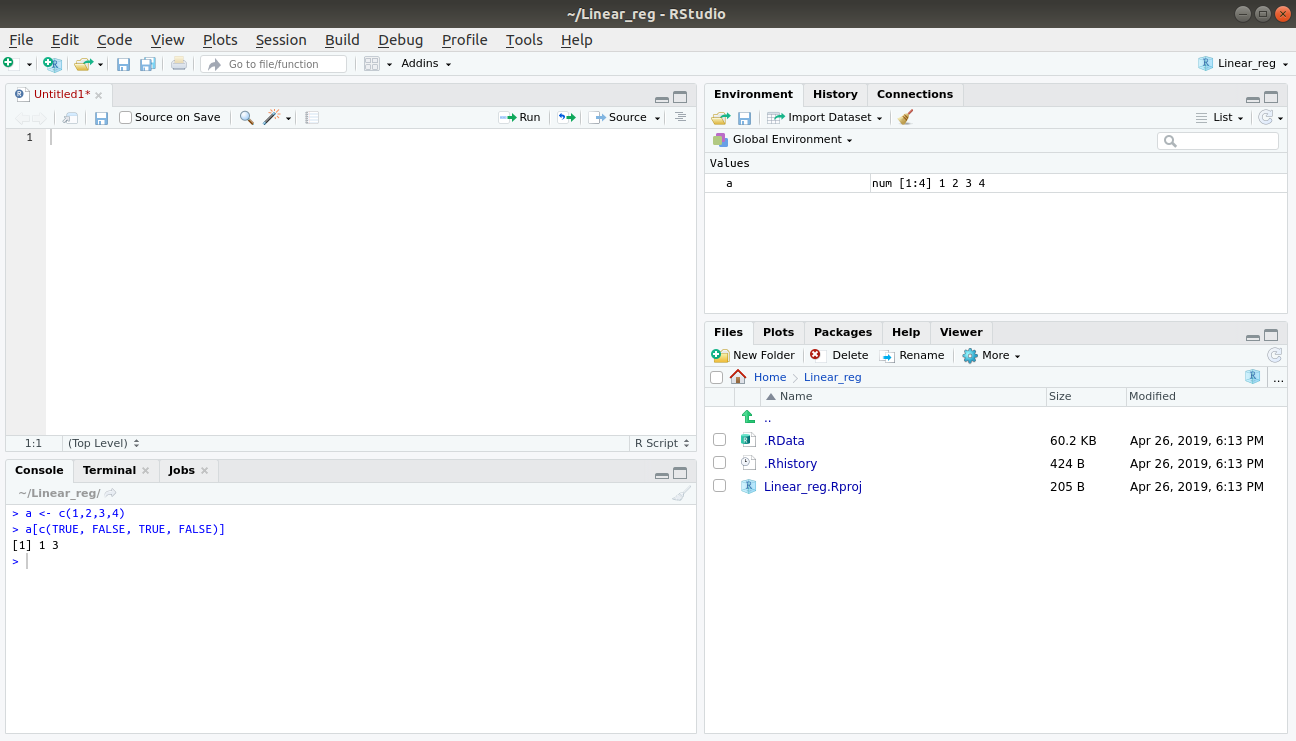
**4.1.4 Sử dụng hàm *seq():***

Có 2 cách sử dụng : seq(first,last,by=number)

seq(first,last,length.out=number)



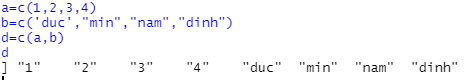
* + 1. **Sử dụng key-value( Mỗi key ứng với 1 value)** 
* Phần tử Logic



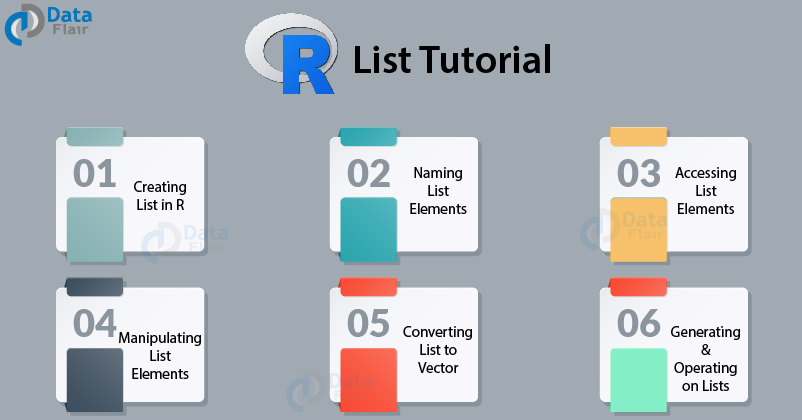
Toán tử trong R Vector :

* + 1. **Combining**





**4.2 LITS**

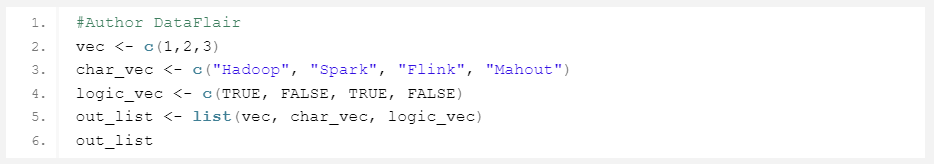


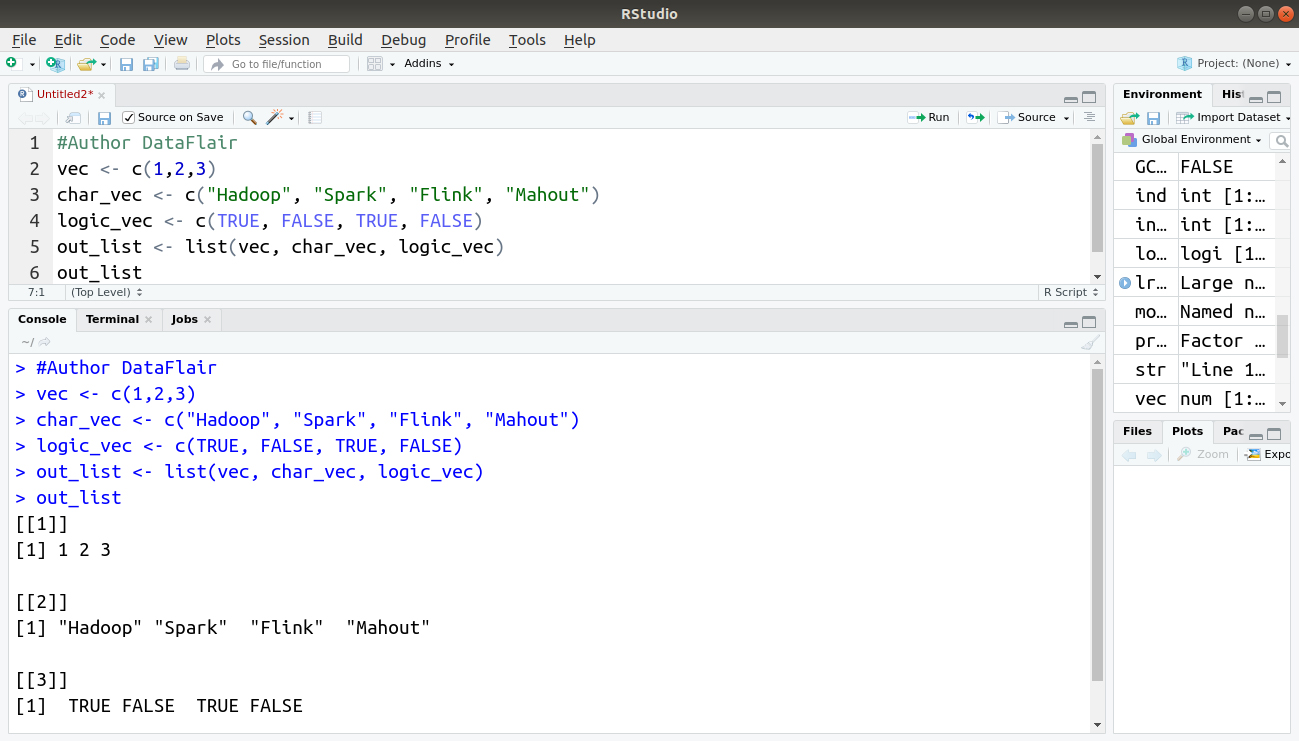
**4.2.1 Lists là gì ?**

Danh sách R là đối tượng chứa các phần tử thuộc các loại khác nhau - như chuỗi, số, vectơ và một danh sách khác bên trong nó. Danh sách R cũng có thể chứa một ma trận hoặc một hàm như các phần tử của nó

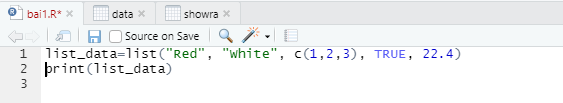
Có thể hiểu Lists chứa nhiều vector

Ví dụ :



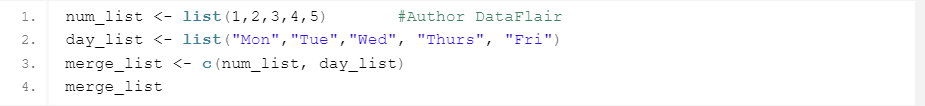


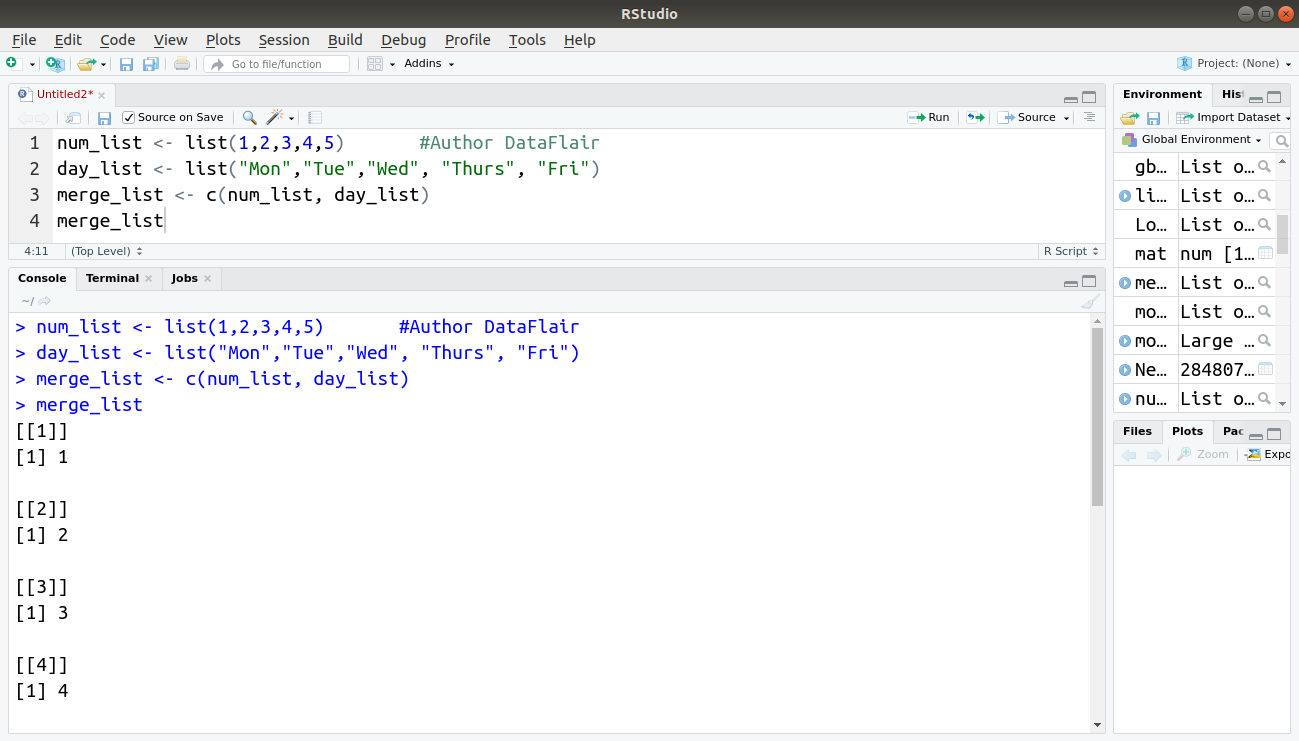
**Cách tạo Lists trong R**





**Cách trộn Lists trong R**





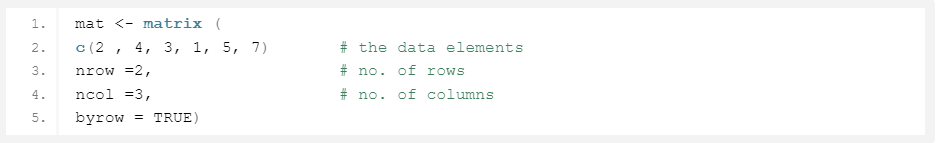
**4.3 Matrix**

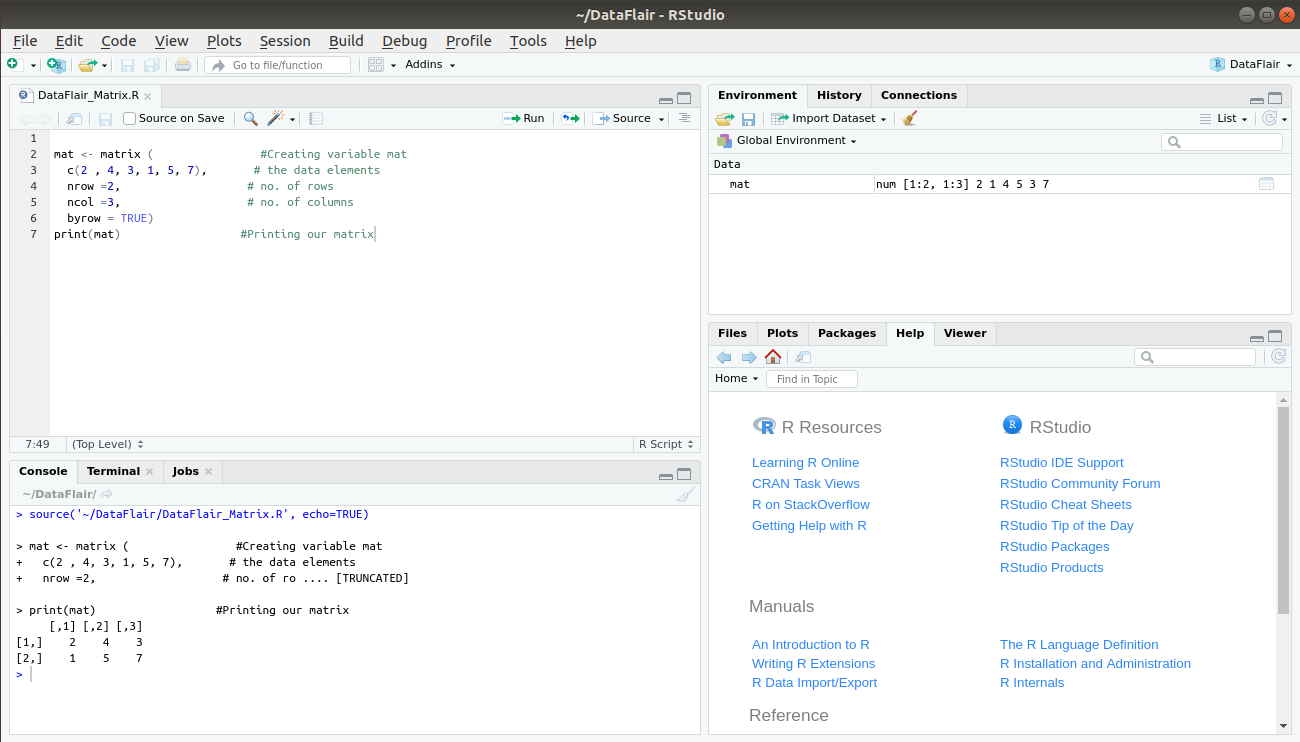
**4.3.1 Matrix là gì ?**

Trong một ma trận, các số được sắp xếp theo một số hàng và cột cố định và thông thường, các số là số thực. Với sự trợ giúp của hàm ma trận, một biểu diễn bộ nhớ của ma trận có thể dễ dàng được tái sử dụng

Tạo ma trận : gồm matrix(vector, nrow=số hàng, ncol=số cột, byrow=TRUE)

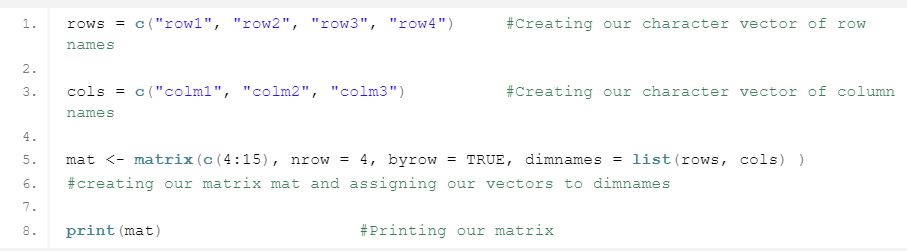
Ví dụ :

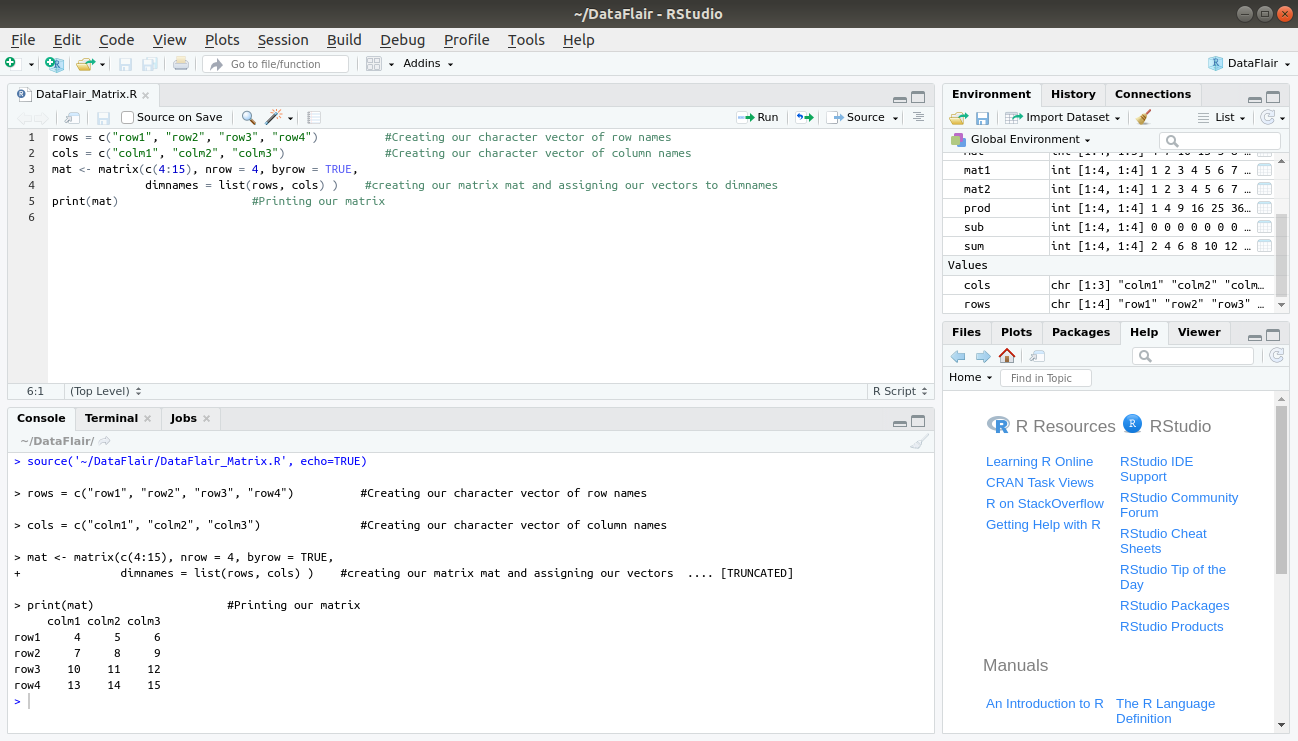




**Định tên cho hàng và cột**

Sử dụng : dinames=list(c(1),c(2))



Output:

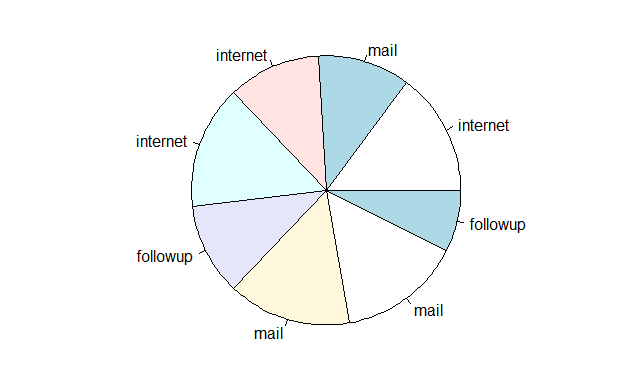
**Graphical Data Analysis**

* 1. **Graphical Data Analysis trong R là gì ?**
* Phần lớn phân tích thống kê dựa trên các kỹ thuật số, chẳng hạn như khoảng tin cậy, kiểm tra giả thuyết, phân tích hồi quy, v.v. Trong nhiều trường hợp, các kỹ thuật này dựa trên các giả định về dữ liệu được sử dụng. Một cách để xác định xem dữ liệu có xác nhận các giả định này hay không là phân tích dữ liệu đồ họa với R, vì một biểu đồ có thể cung cấp nhiều thông tin chuyên sâu về các thuộc tính của tập dữ liệu được vẽ.
* Đồ thị rất hữu ích cho dữ liệu không phải là số, chẳng hạn như màu sắc, hương vị, tên thương hiệu, v.v. Khi các số đo khó hoặc không thể tính toán, đồ thị đóng vai trò quan trọng.
* Điện toán thống kê được thực hiện với mục đích tạo ra đồ họa chất lượng cao
* Plots với biến đơn

***Sử dụng dữ liệu vào từ chương 1***

**Ví dụ : sử dụng piechart để hiển thị tỉ lệ bedroom của từng mode**

pie(data\_age$bedrooms,labels = as.character(data\_age$mode))



**Ứng dụng R vào phân tích bài toán phân nhóm khách hàng**

**Tài liệu tham khảo :** [**https://data-flair.training/blogs/r-data-science-project-customer-segmentation/**](https://data-flair.training/blogs/r-data-science-project-customer-segmentation/)