**TRƯỜNG ĐẠI HỌC MỎ ĐỊA CHẤT**

**BỘ MÔN HỆ THỐNG THÔNG TIN & TRI THỨC**

**TIỂU LUẬN MÔN HỌC**

**Ngôn ngữ lập trình R cho phân tích dữ liệu – 7080232**

**Đề tài: Phân tích “Swiss Dataset”**

**Cán bộ giảng dạy**: *GV, Th.S Vũ Lan Phương*

**SV thực hiện**: *Phan Thanh Huyền*

**MSSV**: *1821050957*

**Lớp**: *DCCTHT63B*

Hà Nội, 01/2022.

***Lời cảm ơn***

Tiểu luận môn học Ngôn ngữ lập trình R cho phân tích dữ liệu với Đề tài “Swiss Fertility and Socioeconomic Indicators (1888)” là thành quả sau quá trình cố gắng không ngừng nghỉ của bản thân em. Qua đây, em xin gửi lời cảm ơn chân thành đến những người đã giúp đỡ em trong thời gian học tập – nghiên cứu tiểu luận vừa qua.

Trước hết, em xin chân thành cảm ơn quý thầy, cô khoa Công nghệ thông tin. Đặc biệt là các thầy cô trong bộ môn Hệ thống thông tin và Tri thức đã tận tình truyền dạy và trang bị cho em những kiến thức nòng cốt trong suốt hành trình ngồi trên ghế giảng đường, làm nền móng giúp em có thể hoàn thành đồ án này một cách trọn vẹn nhất. Tiếp đó, em cũng xin trân trọng gửi lời tri ân đến GV, Th.S Vũ Lan Phương đã dành tâm huyết giúp đỡ, định hướng cho em cách tư duy và làm việc khoa học hơn. Đó là những lời góp ý hết sức quý báu không chỉ trong quá trình thực hiện đồ án này mà còn là hành trang tiếp bước cho em trong quá trình học tập và lập nghiệp sau này.

***Mục lục***

[**Chương I: Tổng quan** 5](#_Toc92412925)

[**I.** **Phân tích dữ liệu là gì?** 5](#_Toc92412926)

[1. Data Analytics là gì? 5](#_Toc92412927)

[2. Các loại phân tích dữ liệu 5](#_Toc92412928)

[3. Ứng dụng của phân tích dữ liệu 7](#_Toc92412929)

[**II.** **Mô tả bài toán** 9](#_Toc92412930)

[1. Thu thập dữ liệu 9](#_Toc92412931)

[2. Mô tả dữ liệu 10](#_Toc92412932)

[**Chương II: Phân tích mô tả** 10](#_Toc92412933)

[**I.** **Khám phá dữ liệu** 10](#_Toc92412934)

[1. Lấy thông tin từ tập dữ liệu 10](#_Toc92412935)

[2. Tải tập dữ liệu swiss được chỉ định sẵn 11](#_Toc92412936)

[3. Kích thước của tập dữ liệu 11](#_Toc92412937)

[4. Hiển thị một phần dữ liệu 11](#_Toc92412938)

[**II.** **Descriptive Analysis** 13](#_Toc92412939)

[1. Summary Statistics 13](#_Toc92412940)

[2. Boxplot 13](#_Toc92412941)

[3. Describe 14](#_Toc92412942)

[4. Correlation 15](#_Toc92412943)

[5. Scatter plot of matrices 16](#_Toc92412944)

[6. Ferility’s Distribution 17](#_Toc92412945)

[7. Education’s Distributions 19](#_Toc92412946)

[8. Agriculture’s Frequency Distribution 20](#_Toc92412947)

[9. Catholic’s Distribution 22](#_Toc92412948)

[10. Examination’s Distribution 23](#_Toc92412949)

[11. Infant.Mortility’s Distribution 24](#_Toc92412950)

[**III.** **Inferential Analysis** 25](#_Toc92412951)

[1. Fertility’s t-test 25](#_Toc92412952)

[2. Education’s t-test 25](#_Toc92412953)

[3. Agriculture’s t-test 26](#_Toc92412954)

[4. Examination’s t-test 26](#_Toc92412955)

[5. Catholic’s t-test 27](#_Toc92412956)

[6. Infant.Mortality’s t-test 27](#_Toc92412957)

[**IV.** **Linear Model** 28](#_Toc92412958)

# **Chương I: Tổng quan**

## **Phân tích dữ liệu là gì?**

Ngày nay, lĩnh vực data analytics dần trở nên quan trọng. Bất kỳ công ty nào đang muốn mở rộng quy mô dù muốn hay không cũng phải cần quan tâm đến nó. Một lượng thông tin khổng lồvẫn đang được thu thập hàng năm và cần phải phân tích cẩn thận để biến tất cả dữ liệu đó thành thông tin chi tiết hữu ích.

Lĩnh vực này ngày càng hấp dẫn hơn nữa khi xuất hiện các chuyên viên data analyst những người kể chuyện bằng con số. Công việc của họ là biến những dữ liệu thô thành chiến lược cho công ty. Một công việc công nghệ đáng mơ ước với mức lương cao, một thị trường việc làm ổn định với đầy những cơ hội làm việc từ xa…

### Data Analytics là gì?

Data analytics là thực hành thu thập và sắp xếp thông tin, tìm kiếm các mẫu và trực quan hóa những phát hiện đó một cách rõ ràng, dễ hiểu.

Hay như **wikipedia** có định nghĩa: “*Phân tích dữ liệu là một quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, thông báo kết luận và hỗ trợ ra quyết định.*

*Phân tích dữ liệu có nhiều khía cạnh và cách tiếp cận, bao gồm các kỹ thuật đa dạng dưới nhiều tên khác nhau và được sử dụng trong các lĩnh vực kinh doanh, khoa học và khoa học xã hội khác nhau.*

*Trong thế giới kinh doanh ngày nay, phân tích dữ liệu đóng vai trò giúp đưa ra quyết định khoa học hơn và giúp doanh nghiệp hoạt động hiệu quả hơn*.”

### Các loại phân tích dữ liệu

#### Text Analysis (Phân tích văn bản)

Text Analysis là một nhánh của Data Mining. Mục đích của nó là tìm kiếm, trích xuất thông tin trong văn bản, chuyển dữ liệu thô thành thông tin kinh doanh. Các công cụ Business Intelligence được sử dụng nhằm đưa ra những quyết định kinh doanh chiến lược. Nhìn chung, nó cung cấp cách để trích xuất và kiểm tra dữ liệu, bắt nguồn từ một mẫu và cuối cùng là giải thích dữ liệu. Hiện nay, dữ liệu văn bản ngày càng nhiều khiến text analysis ngày càng có nhiều ứng dụng trong thực tế. Ví dụ như đối chiếu lý lịch cá nhân, lọc thư rác, phân tích cảm nghĩ, phân loại tài liệu.

#### Statistical Analysis (Phân tích thống kê)

Statistical Analysis thể hiện “Điều gì xảy ra?” bằng cách sử dụng dữ liệu trong quá khứ dưới dạng dashboards. Phân tích thống kê gồm việc thu thập, phân tích, giải thích, trình bày và mô hình hóa dữ liệu. Nó phân tích một tập hợp dữ liệu hoặc một mẫu dữ liệu. Có hai loại phân tích – Descriptive Analysis (phân tích mô tả) và Inferential Analysis (phân tích suy luận).

* Descriptive Analysis (phân tích mô tả): được sử dụng để phân tích và mô tả các dữ liệu đã có sẵn hay dữ liệu lịch sử của doanh nghiệp. Dữ liệu trong quá khứ thường được sử dụng để so sánh và rút ra kết luận. Hai kỹ thuật chính được sử dụng trong loại phân tích này là data aggregation (tổng hợp dữ liệu) và data mining (khai phá dữ liệu). Đầu tiên dữ liệu được thu thập, sau đó được sắp xếp thông qua quá trình tổng hợp dữ liệu. Tiếp theo, các mẫu và ý nghĩa được xác định thông qua quá trình khai phá dữ liệu.
* Inferential Analysis (phân tích suy luận): phân tích mẫu từ dữ liệu hoàn chỉnh. Trong loại phân tích này, bạn có thể tìm thấy các kết luận khác nhau từ cùng một dữ liệu nếu chọn các mẫu khác nhau.

#### Diagnostic Analysis: phân tích chẩn đoán

Đôi khi, có thể có những sai sót trong dữ liệu. Những bất thường này có thể ảnh hưởng tiêu cực đến hoạt động của doanh nghiệp. Mục đích của phân tích chẩn đoán là để chẩn đoán những vấn đề và những lỗi phát sinh này. Hiệu suất và chiến lược của một doanh nghiệp có thể được cải thiện nếu các vấn đề được chẩn đoán và khắc phục sớm.

#### Predictive Analysis: phân tích dự đoán

Như tên gọi của nó, loại phân tích này được sử dụng để dự đoán kết quả trong tương lai. Dựa trên dữ liệu quá khứ hoặc lịch sử, một tập hợp các sự kiện có khả năng xảy ra trong tương lai sẽ được xác định. Phân tích mô tả cũng là một phần của quá trình này, vì nó được sử dụng để tạo ra các dự đoán. Trong loại phân tích này, một lượng lớn dữ liệu được khai thác một cách khéo léo bằng cách sử dụng các mô hình dự đoán. Điều này đảm bảo rằng các dự đoán chính xác có thể được thực hiện.

Ví dụ đơn giản nhất về phân tích dự đoán là nếu năm ngoái bạn mua hai chiếc váy dựa trên khoản tiết kiệm của mình và năm nay lương của bạn tăng gấp đôi nên bạn dự đoán có thể mua được bốn chiếc váy. Nhưng tất nhiên không dễ như thế vì bạn phải suy nghĩ về các trường hợp có thể xảy ra như giá quần áo tăng trong năm nay hoặc có thể thay vì váy bạn muốn mua xe đạp mới, hoặc bạn cần mua nhà.

#### Prescriptive Analysis: phân tích đề xuất

Đây là sự kết hợp của tất cả các loại phân tích khác. Thay vì tập trung vào giám sát dữ liệu, kỹ thuật này nhấn mạnh vào những hiểu biết sâu sắc về dữ liệu từ đó đề xuất những hành động phù hợp và dự đoán kết quả có thể xảy ra. Dữ liệu được thu thập bằng cách sử dụng cả mô hình mô tả và dự đoán, cũng như sự kết hợp giữa toán học và khoa học máy tính.

### Ứng dụng của phân tích dữ liệu

#### Tối ưu hóa marketing

Marketing đã tiến hóa từ một quy trình sáng tạo thành một quy trình phụ thuộc chặt chẽ với dữ liệu. Các tổ chức marketing sử dụng phân tích dữ liệu nhằm xác định kết quả của các chiến dịch và nỗ lực marketing và hướng dẫn quyết định đầu tư cũng như định vị mục tiêu khách hàng. Nghiên cứu nhân khẩu học, xác định phân khúc khách hàng, phân tích kết hợp và các kỹ thuật khác cho phép người làm marketing sử dụng những khối lượng dữ liệu lớn về mua hàng tiêu dùng, khảo sát và nhóm để hiểu và truyền đạt chiến lược marketing

#### Phân tích web

Cho phép những người làm marketing thu thập thông tin về phiên truy cập và tương tác trên các trang web. Google Analytics là một ví dụ điển hình của công cụ miễn phí và phổ thông của mục đích này. Các tương tác này cung cấp cho hệ thống phân tích dữ liệu web những thông tin cần thiết để theo dõi nguồn giới thiệu, tìm kiếm từ khóa, xác định địa chỉ IP, theo dõi hoạt động của khách tới thăm. Với những thông tin này, người làm marketing có thể cải thiện các chiến dịch marketing, nội dung sáng tạo trên website và kiến trúc thông tin của trang web.

Các kỹ thuật phân tích thường sử dụng trong marketing bao gồm mô hình marketing hỗn hợp, phân tích định giá và chiết khấu, tối ưu hóa lực lượng bán hàng và phân tích khách hàng (như xác định phân khúc). Phân tích web và tối ưu hóa trang web và các chiến dịch trực tuyến ngày nay thường kết hợp chặt chẽ với các kỹ thuật phân tích marketing truyền thống. Ảnh hưởng của truyền thông kỹ thuật số cũng tạo ra một số thay đổi về các thuật ngữ, do đó mô hình marketing hỗn hợp cũng thường được tham chiếu tới mô hình phân phối trong ngữ cảnh kỹ thuật số hoặc marketing hỗn hợp.

Các công cụ và kỹ thuật này hỗ trợ cả về các quyết định chiến lược marketing (như dành tổng cộng bao nhiêu ngân sách cho marketing, phân bố ngân sách cho các nhãn hàng và marketing hỗn hợp như thế nào) và về các chiến thuật để định vị khách hàng tiềm năng nhất với thông điệp tối ưu, bằng phương pháp hiệu quả chi phí nhất trong một thời gian lý tưởng.

#### Phân tích dữ liệu con người

Ứng dụng này của phân tích dữ liệu hỗ trợ các công ty quản lý về mặt nhân sự, với mục tiêu là lựa chọn những nhân viên nào để tuyển dụng, khen thưởng hoặc thăng cấp, giao nhiệm vụ gì và các vấn đề nhân sự khác. Ví dụ, một phân tích có thể cho thấy rằng những cá nhân với một kiểu lý lịch nhất định sẽ có nhiều khả năng thành công ở một vị trí cụ thể nào đó, do đó họ là những người phù hợp nhất nên được tuyển dụng. Phân tích nhân sự đang trở nên ngày càng quan trọng để hiểu rõ những hồ sơ với kiểu hành vi nào sẽ thành công hay thất bại. Trong khi phân tích con người được áp dụng cho các nhân viên trong một tổ chức, các kỹ thuật xác định phân khúc khách hàng lại được sử dụng để nghiên cứu hồ sơ khách hàng và phát hiện những khách hàng tiềm năng nhất của thị trường.

#### Phân tích dữ liệu danh mục

Một ứng dụng phổ biến của phân tích dữ liệu kinh doanh là phân tích danh mục. Trong đó, một ngân hàng hoặc tổ chức cho vay có một tập hợp các tài khoản khách hàng với nhiều biến số về giá trị và rủi ro. Các khách hàng này có thể khác nhau về địa vị xã hội (giàu có, trung lưu hoặc nghèo, v.v...), vị trí địa lý, giá trị ròng và các yếu tố khác. Người cho vay phải cân bằng giữa lợi nhuận thu được trên khoản vay với rủi ro vỡ nợ cho từng khoản vay. Câu hỏi đặt ra là làm thế nào để đánh giá được danh mục một cách tổng thể.

Khoản vay ít rủi ro nhất có thể là cho những đối tượng rất giàu có, nhưng có rất ít đối tượng thuộc loại này. Mặt khác, có rất nhiều người nghèo có thể cho vay, tất nhiên là với rủi ro lớn hơn. Một vài số dư cần phải giới hạn nhằm tối ưu hóa lợi nhuận và tối thiểu hóa rủi ro. Giải pháp phân tích dữ liệu có thể kết hợp phân tích chuỗi thời gian với nhiều vấn đề khác để đưa ra quyết định về việc khi nào nên cho vay với từng nhóm phân khúc khách hàng, hoặc quyết định về mức lãi suất cho từng đối tượng trong phân khúc danh mục để bù đắp tổn thất từ toàn thể đối tượng trong danh mục đó.

#### Phân tích dữ liệu rủi ro

Các mô hình dự báo trong ngành ngân hàng được phát triển nhằm đưa ra sự chắc chắn cho chỉ số rủi ro của từng khách hàng riêng lẻ. Chỉ số tín dụng được xây dựng để dự báo hành vi phạm pháp của các cá nhân và được sử dụng rộng rãi để đánh giá mức tín dụng xứng đáng của người nộp hồ sơ vay vốn.

Bên cạnh đó, phân tích rủi ro được thực hiện trong khoa học và lĩnh vực bảo hiểm. Nó cũng được dùng rộng rãi trong các tổ chức tài chính như các công ty về cổng thanh toán trực tuyến để phân tích xem một giao dịch là có thực hay gian lận bằng việc sử dụng lịch sử giao dịch của khách hàng. Ứng dụng này được dùng rộng rãi hơn trong mua hàng thanh toán bằng thẻ tín dụng, ví dụ khi có sự tăng đột biến trong khối lượng giao dịch của một khách hàng, khách hàng đó sẽ nhận được cuộc gọi để xác nhận giao dịch đó có phải được khởi tạo bởi họ hay không. Ứng dụng này sẽ giúp giảm thiểu tổn thất trong những trường hợp kể trên.

#### Phân tích dữ liệu kỹ thuật số

Phân tích dữ liệu kỹ thuật số là một nhóm các hoạt động kinh doanh và kỹ thuật nhằm xác định, khởi tạo, thu thập, xác minh hay chuyển đổi dữ liệu kỹ thuật số thành báo cáo, nghiên cứu, phân tích, đề xuất, tối ưu hóa, dự báo hoặc tự động hóa. Ứng dụng này cũng bao gồm SEO (Search Engine Optimization - Tối ưu hóa công cụ tìm kiếm) trong đó các tìm kiếm từ khóa được theo dõi và trở thành dữ liệu được sử dụng cho mục đích marketing.

Thậm chí các banner quảng cáo và lần nhấp chuột cũng thuộc phân tích dữ liệu kỹ thuật số. Ngày càng có nhiều công ty về marketing và thương hiệu phụ thuộc vào phân tích dữ liệu kỹ thuật số cho công việc marketing kỹ thuật số của họ, trong đó MORI (Marketing Return On Investment - lợi nhuận marketing từ đầu tư) là một chỉ số hoạt động quan trọng.

#### Phân tích dữ liệu an ninh

Phân tích dữ liệu an ninh có liên quan đến công nghệ thông tin, nhằm thu thập và phân tích các sự kiện an ninh để tìm ra những yếu tố nào mang tới rủi ro lớn nhất. Sản phẩm trong lĩnh vực này bao gồm quản lý bảo mật thông tin - sự kiện và phân tích dữ liệu hành vi người dùng.

#### Phân tích dữ liệu phần mềm

Phân tích dữ liệu phần mềm là quá trình thu thập và phân tích thông tin về cách thức một phần mềm được sản xuất và sử dụng như thế nào.

## **Mô tả bài toán**

### Thu thập dữ liệu

Một số tệp dữ liệu mẫu được tích hợp sẵn trong package (datasets-package) thường được dùng để demo, luyện tập, thực hành và tăng khả sử dụng ngôn ngữ R. Các tệp dữ liệu trong package đều có thể sử dụng miễn phí, được đóng góp từ R core team và các cộng tác viên trên toàn thế giới.

### Mô tả dữ liệu

Tên dataset: **swiss -** Swiss Fertility and Socioeconomic Indicators (1888) Data

Tiêu chuẩn của thước đo sinh sản và các chỉ số kinh tế xã hội cho 47 tỉnh nói tiếng Pháp của Thụy Sĩ vào khoảng năm 1888.

Mỗi dòng ứng với các thông tin của mỗi tỉnh, bao gồm 6 thuộc tính:

|  |  |  |
| --- | --- | --- |
| **[1]** | **Fertility** | % thước đo mức sinh tiêu chuẩn hóa chung |
| **[2]** | **Agriculture** | % nam giới tham gia vào nông nghiệp với tư cách là nghề nghiệp |
| **[3]** | **Examination** | % người đi nghĩa vụ quân sự nhận được điểm cao nhất trong kỳ thi quân đội |
| **[4]** | **Education** | % giáo dục ngoài trường tiểu học cho người đi nghĩa vụ quân sự. |
| **[5]** | **Catholic** | % người theo đạo công giáo |
| **[6]** | **Infant.Mortality** | % tử vong là trẻ sơ sinh sống dưới 1 năm |

# **Chương II: Phân tích mô tả**

## **Khám phá dữ liệu**

### Lấy thông tin từ tập dữ liệu

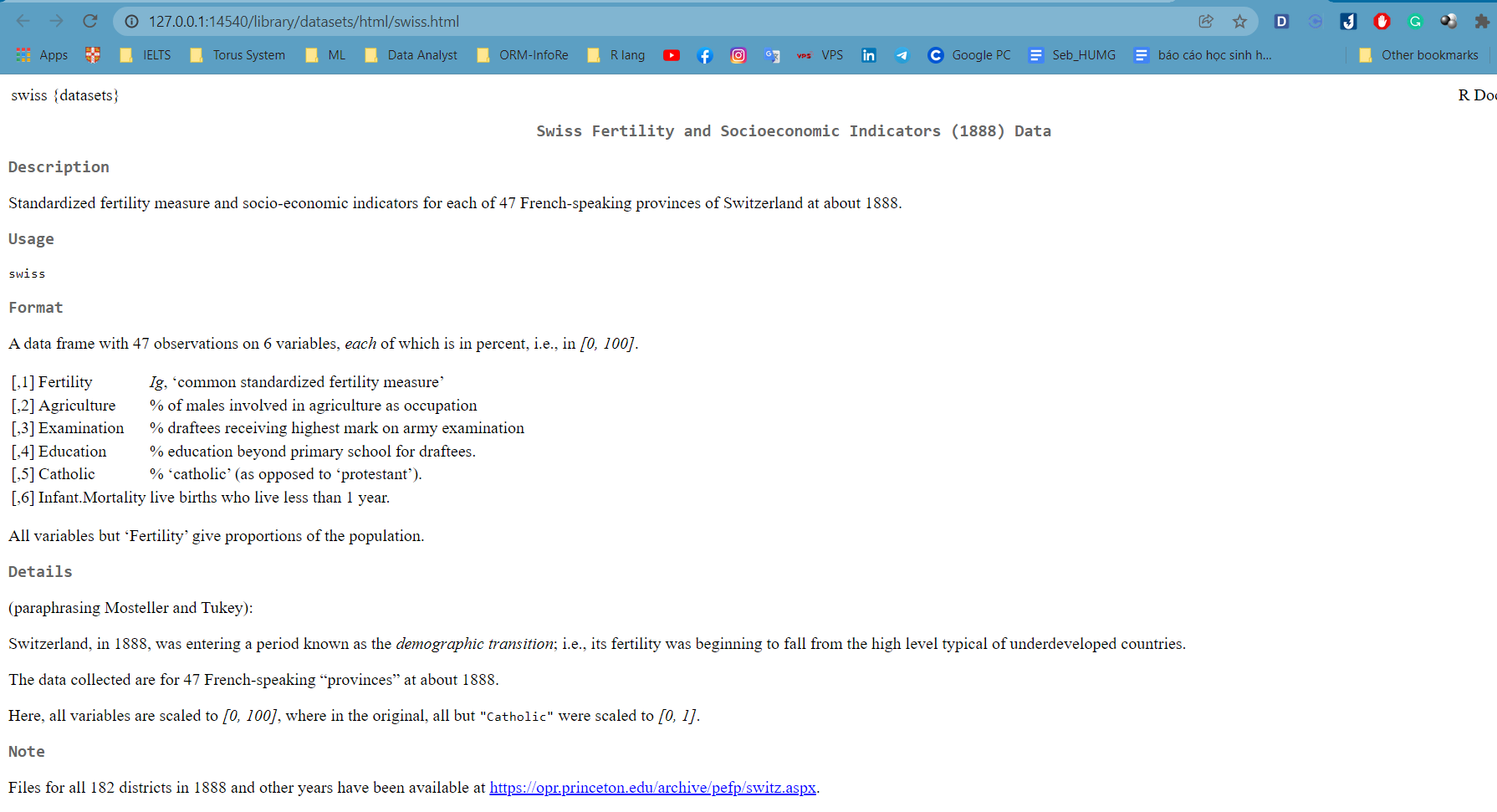
Lấy thông tin trên tập dữ liệu bằng lệnh ?dataset, thông tin sẽ được chuyển tới trang thông tin của tập dữ liệu Swiss như hình dưới:

#get information on the dataset

?swiss

> ?swiss

starting httpd help server ... done



### Tải tập dữ liệu swiss được chỉ định sẵn

#load data

data(swiss)

### Kích thước của tập dữ liệu

Hàm dim() sẽ trả về kích thước của tập dữ liệu, kết quả được trả về là tập dữ liệu chứa 47 hàng và 6 thuộc tính:

#returns the dimension

dim(swiss)

> dim(swiss)

[1] 47 6

### Hiển thị một phần dữ liệu

Với 1 số lệnh bên dưới sẽ hiển thị ra từng phần của dữ liệu mà chúng ta muốn xem:

> head(swiss)

Fertility Agriculture Examination Education Catholic

Courtelary 80.2 17.0 15 12 9.96

Delemont 83.1 45.1 6 9 84.84

Franches-Mnt 92.5 39.7 5 5 93.40

Moutier 85.8 36.5 12 7 33.77

Neuveville 76.9 43.5 17 15 5.16

Porrentruy 76.1 35.3 9 7 90.57

Infant.Mortality

Courtelary 22.2

Delemont 22.2

Franches-Mnt 20.2

Moutier 20.3

Neuveville 20.6

Porrentruy 26.6

Lệnh head() sẽ hiển thị 6 dòng đầu tiên của tập dữ liệu như bên trên.

Tương tự, ta có lệnh tail() để hiển thị 6 dòng cuối cùng của tập dữ liệu:

> tail(swiss)

Fertility Agriculture Examination Education Catholic

Neuchatel 64.4 17.6 35 32 16.92

Val de Ruz 77.6 37.6 15 7 4.97

ValdeTravers 67.6 18.7 25 7 8.65

V. De Geneve 35.0 1.2 37 53 42.34

Rive Droite 44.7 46.6 16 29 50.43

Rive Gauche 42.8 27.7 22 29 58.33

Infant.Mortality

Neuchatel 23.0

Val de Ruz 20.0

ValdeTravers 19.5

V. De Geneve 18.0

Rive Droite 18.2

Rive Gauche 19.3

Hiển thị tên của 47 tỉnh/thành phố ở trong tập dữ liệu

> row.names(swiss)

[1] "Courtelary" "Delemont" "Franches-Mnt" "Moutier"

[5] "Neuveville" "Porrentruy" "Broye" "Glane"

[9] "Gruyere" "Sarine" "Veveyse" "Aigle"

[13] "Aubonne" "Avenches" "Cossonay" "Echallens"

[17] "Grandson" "Lausanne" "La Vallee" "Lavaux"

[21] "Morges" "Moudon" "Nyone" "Orbe"

[25] "Oron" "Payerne" "Paysd'enhaut" "Rolle"

[29] "Vevey" "Yverdon" "Conthey" "Entremont"

[33] "Herens" "Martigwy" "Monthey" "St Maurice"

[37] "Sierre" "Sion" "Boudry" "La Chauxdfnd"

[41] "Le Locle" "Neuchatel" "Val de Ruz" "ValdeTravers"

[45] "V. De Geneve" "Rive Droite" "Rive Gauche"

## **Descriptive Analysis**

### Summary Statistics

> summary(swiss)

Fertility Agriculture Examination Education

Min. :35.00 Min. : 1.20 Min. : 3.00 Min. : 1.00

1st Qu.:64.70 1st Qu.:35.90 1st Qu.:12.00 1st Qu.: 6.00

Median :70.40 Median :54.10 Median :16.00 Median : 8.00

Mean :70.14 Mean :50.66 Mean :16.49 Mean :10.98

3rd Qu.:78.45 3rd Qu.:67.65 3rd Qu.:22.00 3rd Qu.:12.00

Max. :92.50 Max. :89.70 Max. :37.00 Max. :53.00

Catholic Infant.Mortality

Min. : 2.150 Min. :10.80

1st Qu.: 5.195 1st Qu.:18.15

Median : 15.140 Median :20.00

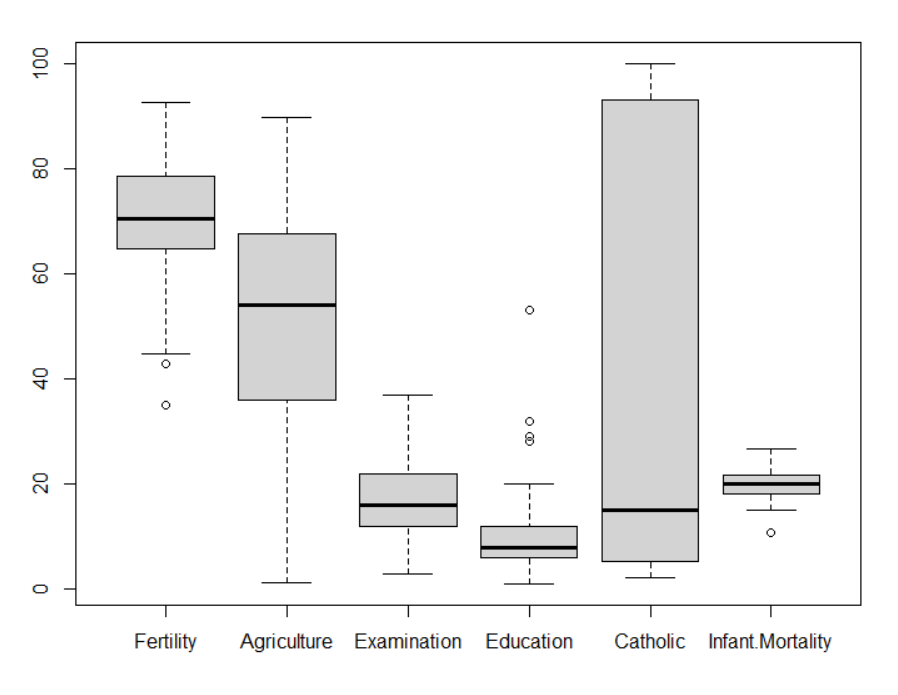
Mean : 41.144 Mean :19.94

3rd Qu.: 93.125 3rd Qu.:21.70

Max. :100.000 Max. :26.60

### Boxplot

> boxplot(swiss)



Fertility: mean và median không quá chênh lệch; các giá trị nằm trong khoảng 35 đến 92.5; có 2 outliers nằm ngoài phạm vi cho phép

Q1 = 64.7 với 25% đối tượng có chỉ số mức sinh nhỏ hơn 64.7

Q2 = 70.4 với 50% bé (hơn) bằng 70.4

Q3 = 78.45 với 25% đối tượng lớn hơn 78.45

* IQR Fertility là 13.75, có 50% đối tượng có chỉ số sinh đẻ nằm trong khoảng từ 64.7 đến 78.45 với biên độ giao động là 13.75

Agriculture: Các giá trị nằm trong phạm vi từ 1.2 đến 89.7, không có outliers.

Q1 = 35.9, Q2 = 54.1, Q3 = 67.65

* IQR Agriculture là 31.75, có 50% đối tượng là nam giới làm nông nghiệp nằm trong khoảng từ 35.9 đến 67.65 với biên độ giao động là 31.75

Examination: Các giá trị nằm trong phạm vi từ 3 đến 37, không có outliers.

Q1 = 12, Q2 = 16, Q3 = 22

* IQR Examination là 10, có 50% đối tượng nằm trong khoảng từ 12 tới 22

Education: Các giá trị nằm trong phạm vi từ 1 đến 53, có 4 outliers nằm ngoài phạm vi cho phép.

Q1 = 6, Q2 = 8, Q3 = 12

* IQR Education là 6, có 50% đối tượng nằm trong khoảng từ 6 đến 12

Catholic: Các giá trị nằm trải dài trong phạm vi từ 2.15 đến 100

Q1 = 5.195, Q2 = 15.14, Q3 = 93.125

* IQR Catholic là 87.93, biên độ giao động trải rộng

Infant.Mortality: Các giá trị nằm trong phạm vi từ 10.8 đến 26.6, có duy nhất 1 outlier

Q1 = 18.15, Q2 = 20, Q3 = 21.7

* IQR Infant.Mortality là 3.55, các giá trị nằm tập trung ở 1 khoảng biên độ nhỏ

### Describe

> library(psych)

Warning message:

package ‘psych’ was built under R version 4.0.5

> describe(swiss)

vars n mean sd median trimmed mad min max range

Fertility 1 47 70.14 12.49 70.40 70.66 10.23 35.00 92.5 57.50

Agriculture 2 47 50.66 22.71 54.10 51.16 23.87 1.20 89.7 88.50

Examination 3 47 16.49 7.98 16.00 16.08 7.41 3.00 37.0 34.00

Education 4 47 10.98 9.62 8.00 9.38 5.93 1.00 53.0 52.00

Catholic 5 47 41.14 41.70 15.14 39.12 18.65 2.15 100.0 97.85

Infant.Mortality 6 47 19.94 2.91 20.00 19.98 2.82 10.80 26.6 15.80

skew kurtosis se

Fertility -0.46 0.26 1.82

Agriculture -0.32 -0.89 3.31

Examination 0.45 -0.14 1.16

Education 2.27 6.14 1.40

Catholic 0.48 -1.67 6.08

Infant.Mortality -0.33 0.78 0.42

### Correlation

> cor(swiss)

Fertility Agriculture Examination Education Catholic

Fertility 1.0000000 0.35307918 -0.6458827 -0.66378886 0.4636847

Agriculture 0.3530792 1.00000000 -0.6865422 -0.63952252 0.4010951

Examination -0.6458827 -0.68654221 1.0000000 0.69841530 -0.5727418

Education -0.6637889 -0.63952252 0.6984153 1.00000000 -0.1538589

Catholic 0.4636847 0.40109505 -0.5727418 -0.15385892 1.0000000

Infant.Mortality 0.4165560 -0.06085861 -0.1140216 -0.09932185 0.1754959

Infant.Mortality

Fertility 0.41655603

Agriculture -0.06085861

Examination -0.11402160

Education -0.09932185

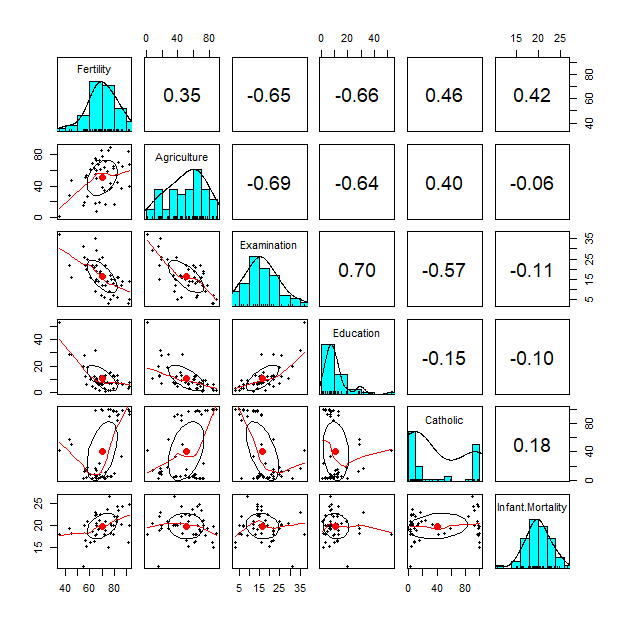
Catholic 0.17549591

Infant.Mortality 1.00000000

### Scatter plot of matrices

> library(psych)

> pairs.panels(swiss)

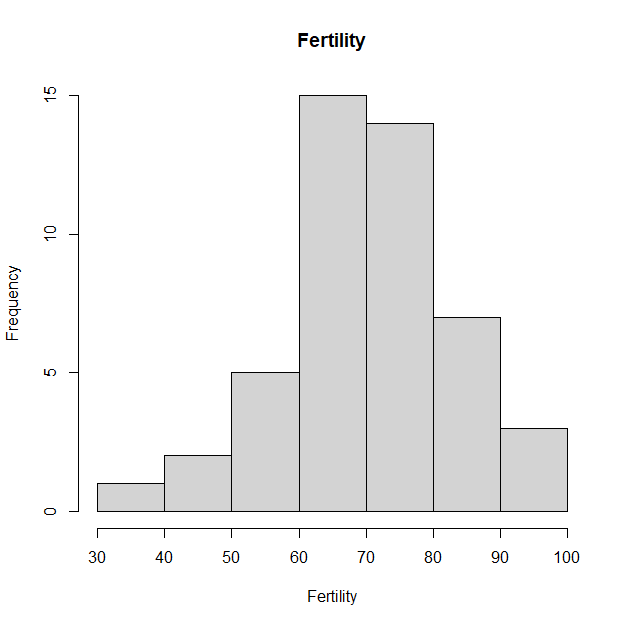


Có thể thấy, giữ Examination và Education có mối tương quan khá lớn, còn lại thì tương quan nhẹ và không liên quan tới nhau, ví dụ như Fertility và Agriculture có mối tương quan nhẹ.

Xét về biểu đồ các biến thì Fertility, Agiculture, Examination, Infant.Mortality có phân phối gần chuẩn; Education bị lệch dương (dữ liệu phân bố nhiều ở phần bên trái); còn Catholic có phân bố không đồng đều, dữ liệu phân bố ở 2 điểm cách xa nhau nhất.

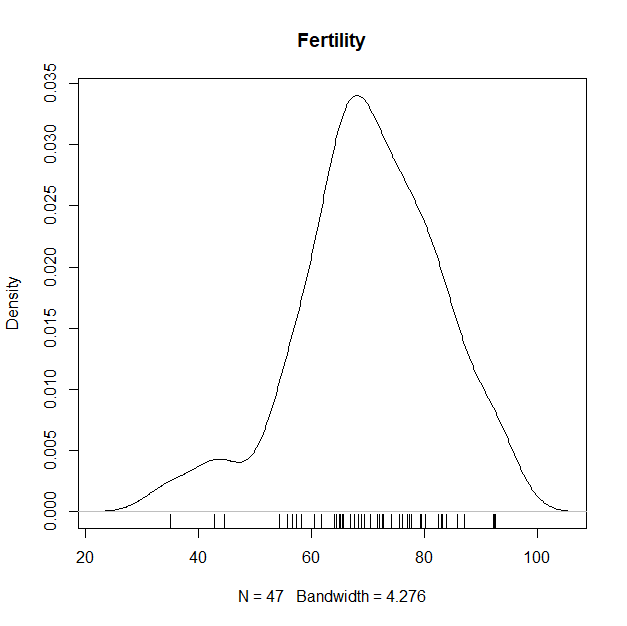
### Ferility’s Distribution

> hist(swiss$Fertility,main="Fertility",xlab="Fertility")



> plot(density(swiss$Fertility),main="Fertility")

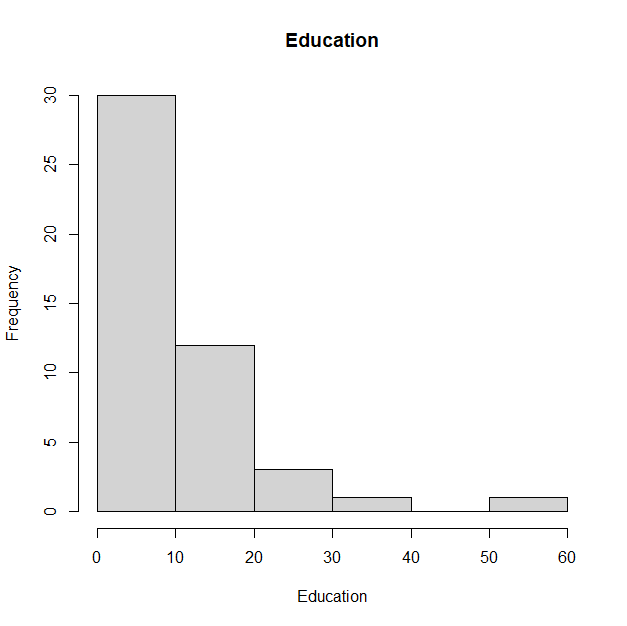
> rug(swiss$Fertility)



Tỉ lệ sinh sản nằm chủ yếu từ 60-90%, biểu đồ có xu hướng lệch phải không đáng kể.

### Education’s Distributions

> hist(swiss$Education,main="Education", xlab = "Education")

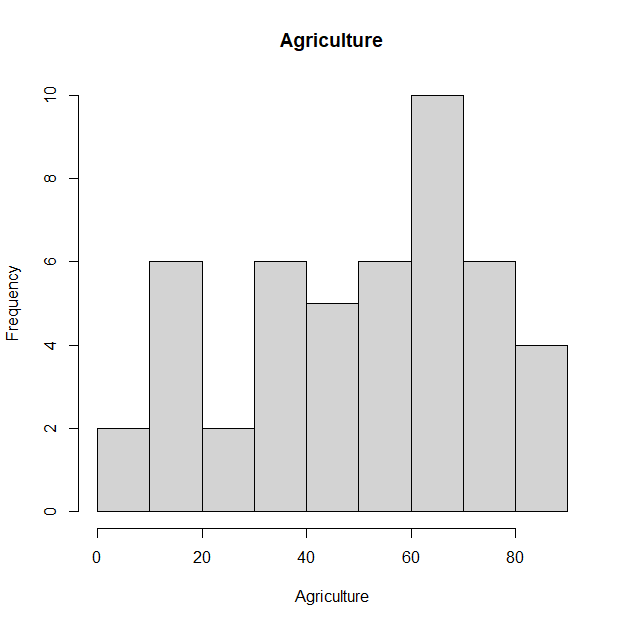


Dữ liệu lệch sang bên trái cho thấy rằng phần lớn các thành phố có dưới 20% số người được giáo dục sau tiểu học, con số này khá khiêm tốn.

### Agriculture’s Frequency Distribution

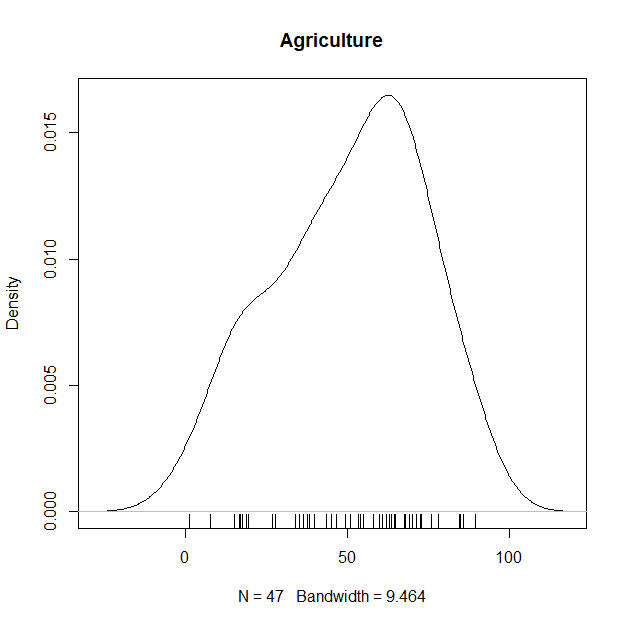
>hist(swiss$Agriculture,main="Agriculture", xlab = "Agriculture")

Đối với biến Agriculture, lúc nhìn vào bảng phân bố tần suất rất khó để xác định vấn đề, em đã thêm biểu đồ mật độ để dễ dàng quan sát các giá trị của biến Agriculture



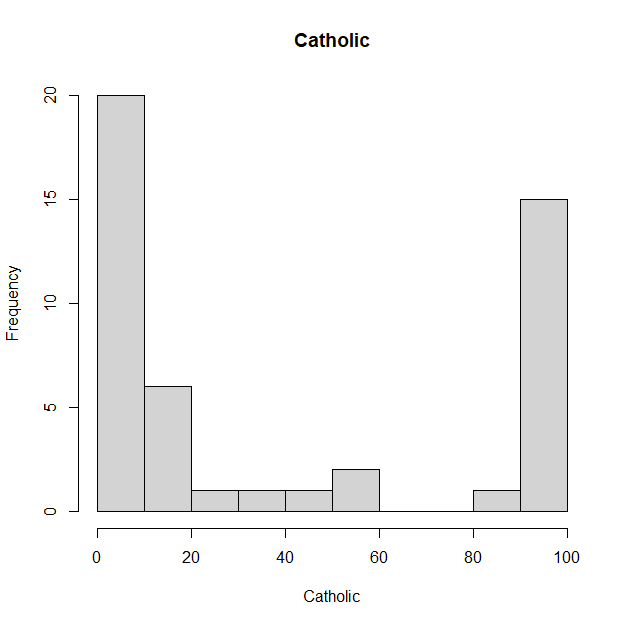
> plot(density(swiss$Agriculture),main="Agriculture")

> rug(swiss$Agriculture)



Có thể thấy, rõ ràng dữ liệu của chúng ta phân bố có phần hơi lệch phải.

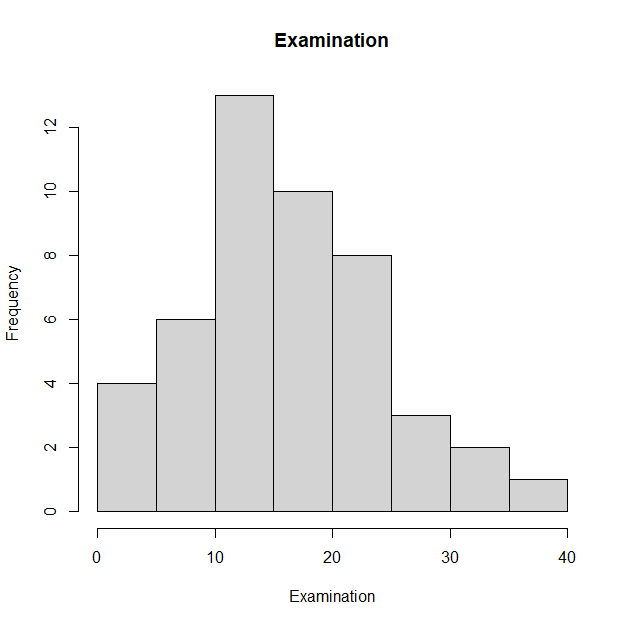
### Catholic’s Distribution



Dữ liệu phân phối ở 2 đầu cực, cho thấy tỉ lệ theo đạo công giáo từng vùng có sự chênh lệch khá lớn.

### Examination’s Distribution

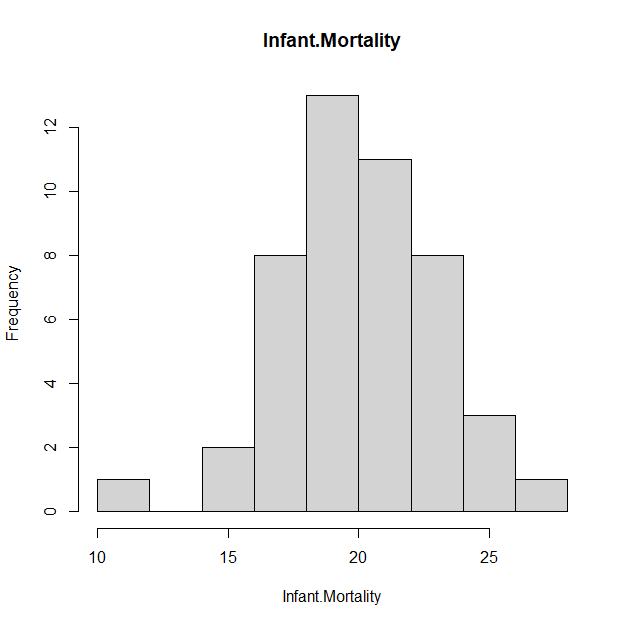
> hist(swiss$Examination,main="Examination", xlab = "Examination")



Dữ liệu bị lệch trái cho thấy ít thành phố có tỷ lệ phần trăm cao những người có điểm thi cao. Phần lớn người đạt điểm cao chỉ chiếm ở 10-20%

### Infant.Mortility’s Distribution

>hist(swiss$Infant.Mortality,main="Infant.Mortality", xlab = "Infant.Mortality")



Có vẻ như tỷ lệ tử vong ở trẻ sơ sinh được phân bổ khá chuẩn với một số tỷ lệ nhỏ hơn ở đuôi bên trái cho thấy một số lượng nhỏ các thành phố có tỷ lệ trẻ sơ sinh thấp.

## **Inferential Analysis**

### Fertility’s t-test

> t.test(swiss$Fertility)

One Sample t-test

data: swiss$Fertility

t = 38.495, df = 46, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

66.47485 73.81025

sample estimates:

mean of x

70.14255

***H0: µ = k***

***H1: µ ≠ k***

Với khoảng tin cậy 95% trung bình tổng thể về mức sinh sản nằm trong khoảng ***k*** từ 66.5% đến 73.8%. Trung bình thực của tỉ lệ mức sinh là 70, nằm trong khoảng tin cậy vì vậy chúng ta bác bỏ giả thuyết ***H1***, chấp nhận giả thuyết ***H0***.

### Education’s t-test

> t.test(swiss$Education)

One Sample t-test

data: swiss$Education

t = 7.8277, df = 46, p-value = 5.314e-10

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

8.155534 13.801913

sample estimates:

mean of x

10.97872

***H0: µ = k***

***H1: µ ≠ k***

Với khoảng tin cậy 95% trung bình tổng thể về giáo dục ngoài tiểu học nằm trong khoảng ***k*** từ 8.15% đến 13.8%. Trung bình thực của tỉ lệ giáo dục ngoài tiểu học là 10.97, nằm trong khoảng tin cậy vì vậy chúng ta bác bỏ giả thuyết ***H1***, chấp nhận giả thuyết ***H0***.

### Agriculture’s t-test

> t.test(swiss$Agriculture)

One Sample t-test

data: swiss$Agriculture

t = 15.292, df = 46, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

43.99131 57.32784

sample estimates:

mean of x

50.65957

***H0: µ = k***

***H1: µ ≠ k***

Với khoảng tin cậy 95% trung bình tổng thể về giáo dục ngoài tiểu học nằm trong khoảng ***k*** từ 44% đến 57.32%. Trung bình thực của tỉ lệ giáo dục ngoài tiểu học là 50.66%, nằm trong khoảng tin cậy vì vậy chúng ta bác bỏ giả thuyết ***H1***, chấp nhận giả thuyết ***H0***.

### Examination’s t-test

> t.test(swiss$Examination)

One Sample t-test

data: swiss$Examination

t = 14.17, df = 46, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

14.14697 18.83176

sample estimates:

mean of x

16.48936

***H0: µ = k***

***H1: µ ≠ k***

Với khoảng tin cậy 95% trung bình tổng thể về giáo dục ngoài tiểu học nằm trong khoảng ***k*** từ 14.14% đến 18.83%. Trung bình thực của tỉ lệ giáo dục ngoài tiểu học là 16.48%, nằm trong khoảng tin cậy vì vậy chúng ta bác bỏ giả thuyết ***H1***, chấp nhận giả thuyết ***H0***.

### Catholic’s t-test

> t.test(swiss$Catholic)

One Sample t-test

data: swiss$Catholic

t = 6.7634, df = 46, p-value = 2.064e-08

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

28.89883 53.38883

sample estimates:

mean of x

41.14383

***H0: µ = k***

***H1: µ ≠ k***

Với khoảng tin cậy 95% trung bình tổng thể về giáo dục ngoài tiểu học nằm trong khoảng ***k*** từ 28.89% đến 53.38%. Trung bình thực của tỉ lệ giáo dục ngoài tiểu học là 41.14%, nằm trong khoảng tin cậy vì vậy chúng ta bác bỏ giả thuyết ***H1***, chấp nhận giả thuyết ***H0***.

### Infant.Mortality’s t-test

> t.test(swiss$Infant.Mortality)

One Sample t-test

data: swiss$Infant.Mortality

t = 46.939, df = 46, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

19.08735 20.79775

sample estimates:

mean of x

19.94255

***H0: µ = k***

***H1: µ ≠ k***

Với khoảng tin cậy 95% trung bình tổng thể về giáo dục ngoài tiểu học nằm trong khoảng ***k*** từ 19.08% đến 20.79%. Trung bình thực của tỉ lệ giáo dục ngoài tiểu học là 19.9%, nằm trong khoảng tin cậy vì vậy chúng ta bác bỏ giả thuyết ***H1***, chấp nhận giả thuyết ***H0***.

## **Linear Model**

> model <- lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, swiss)

> summary(model)

Call:

lm(formula = Fertility ~ Agriculture + Examination + Education +

Catholic + Infant.Mortality, data = swiss)

Residuals:

Min 1Q Median 3Q Max

-15.2743 -5.2617 0.5032 4.1198 15.3213

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 66.91518 10.70604 6.250 1.91e-07 \*\*\*

Agriculture -0.17211 0.07030 -2.448 0.01873 \*

Examination -0.25801 0.25388 -1.016 0.31546

Education -0.87094 0.18303 -4.758 2.43e-05 \*\*\*

Catholic 0.10412 0.03526 2.953 0.00519 \*\*

Infant.Mortality 1.07705 0.38172 2.822 0.00734 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom

Multiple R-squared: 0.7067, Adjusted R-squared: 0.671

F-statistic: 19.76 on 5 and 41 DF, p-value: 5.594e-10