**ỨNG DỤNG**

**PHÂN TÍCH DỮ LIỆU THÔNG MINH**

**BÁO CÁO**

**ĐỀ TÀI THỰC HÀNH CÁ NHÂN**

Họ và tên : **Vương Thế Khang**

MSSV : **19120539**

C:\Users\tdqua_000\Dropbox\SS-Slides\DeCuong-CDIO\Template CDIO v4.2\Templates\Hinh anh\LogoTruong.png

Bộ môn Khoa học máy tính

Khoa Công nghệ thông tin

Đại học Khoa học tự nhiên TP HCM

**MỤC LỤC**

[1 Thu thập dữ liệu 3](#_Toc153397261)

[2 Khám phá dữ liệu 4](#_Toc153397262)

[**2.1** **Cấu trúc dữ liệu** 4](#_Toc153397263)

[**2.2** **Chất lượng bộ dữ liệu** 6](#_Toc153397264)

[**a.** **Phân bố giá trị của các cột có kiểu dữ liệu dạng số** 6](#_Toc153397265)

[**b.** **Phân bố giá trị của các cột có kiểu dữ liệu dạng phân loại** 7](#_Toc153397266)

[3 Trực quan hóa dữ liệu 10](#_Toc153397267)

[**3.1** **Độ tuổi trung bình của các tỷ phú trên thế giới là bao nhiêu?** 10](#_Toc153397268)

[**3.2** **Có nhiều tỷ phú xuất thân tự kinh doanh giữa nam và nữ?** 11](#_Toc153397269)

[**3.3** **Hiện nay các tỷ phú tập trung nhiều ở các quốc gia nào?** 13](#_Toc153397270)

[**3.4** **Lĩnh vực nào có nhiều nhà tỷ phú đầu tư nhất?** 14](#_Toc153397271)

[**3.5** **Có sự khác nhau về lĩnh vực kinh doanh giữa tỷ phú nam và nữ không?** 15](#_Toc153397272)

[**3.6** **Giá trị tài sản ròng của các tỷ phú được phân bố như thế nào theo quốc tịch?** 16](#_Toc153397273)

[4 Rút ra kết luận 19](#_Toc153397274)

[5 Tài liệu tham khảo 20](#_Toc153397275)

# Thu thập dữ liệu

Bộ dữ liệu được cung cấp dùng để phân tích ra và rút ra các kết luận dựa vào các số liệu thống kê về các tỷ phú trên thế giới. Thời gian dữ liệu được thu thập là 2023/04/04.

Đây là mô tả về bộ dữ liệu trên Kaggle:

*“Bộ dữ liệu này chứa các số liệu thống kê về các tỷ phú trên thế giới, bao gồm thông tin về doanh nghiệp, lĩnh vực kinh doanh và các thông tin cá nhân của họ. Điều này cung cấp một cái nhìn sâu sắc về sự phân bố tài sản, lĩnh vực kinh doanh và nhân khẩu học của các tỷ phú trên toàn thế giới.”*

Thông tin về dữ liệu:

* File dữ liệu: “Billionaires Statistics Dataset.csv”
* [Link Kaggle](https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset/data)

# Khám phá dữ liệu

Đầu tiên, ta sẽ xem thử bộ dữ liệu trông như thế nào. Đọc dữ liệu từ file "Billionaires Statistics Dataset.csv" và lưu vào dataframe billion\_df.

Ảnh có chứa ảnh chụp màn hình, văn bản

Mô tả được tạo tự động

* 1. **Cấu trúc dữ liệu**

Ta thấy được:

* Dữ liệu có kích thước **2,640 dòng x 35 cột**.
* Mỗi dòng chứa thông tin về một tỷ phú trong bảng xếp hạng.
* Không có dòng nào bị lặp và không có ý nghĩa khác nhau giữa các dòng.

Mỗi cột của tập dữ liệu billion\_df có ý nghĩa như sau:

|  |  |
| --- | --- |
| **Cột** | **Ý nghĩa** |
| rank | xếp hạng của mỗi tỷ phú theo mức độ giàu có |
| finalWorth | giá trị ròng cuối cùng của tỷ phú (triệu đô la Mỹ) |
| category | danh mục hoặc lĩnh vực doanh nghiệp của tỷ phú đang hoạt động |
| personName | tên đầy đủ của tỷ phú |
| age | tuổi của tỷ phú |
| country | quốc gia nơi tỷ phú đang sinh sống |
| city | thành phố nơi tỷ phú đang sinh sống |
| source | nguồn lợi nhuận của tỷ phú (công ty, doanh nghiệp, …) |
| industries | các lĩnh vực kinh doanh hoặc gắn liền với lợi ích kinh doanh của tỷ phú |
| countryOfCitizenship | quốc tịch của tỷ phú |
| organization | tên của tổ chức hoặc công ty liên kết với tỷ phú |
| selfMade | cho biết tỷ phú có tự lập kinh doanh không |
| status | cho biết trạng thái của tỷ phú là tự lập kinh doanh (nhà sáng lập/doanh nhân) hoặc được thừa kế tài sản hoặc không kiếm được |
| gender | giới tính của tỷ phú |
| birthData | ngày tháng năm sinh của tỷ phú |
| lastName | họ của tỷ phú |
| firstName | tên của tỷ phú |
| title | danh hiệu của tỷ phú |
| date | ngày thu thập dữ liệu |
| state | bang nơi tỷ phú đang sinh sống |
| residenceStateRegion | khu vực hoặc tiểu bang cư trú của tỷ phú |
| birthYear | năm sinh của tỷ phú |
| birthMonth | tháng sinh của tỷ phú |
| birthDay | ngày sinh của tỷ phú |
| cpi\_country | chỉ số giá tiêu dùng (CPI) của quốc gia tỷ phú |
| cpi\_change\_country | CPI thay đổi của quốc gia tỷ phú |
| gdp\_country | tổng sản phẩm trong nước (GDP) của quốc gia tỷ phú |
| gross\_tertiary\_  education\_enrollment | tuyển sinh vào giáo dục đại học ở quốc gia tỷ phú |
| gross\_primary\_  education\_enrollment\_  country | tuyển sinh vào giáo dục tiểu học ở quốc gia tỷ phú |
| life\_expectancy\_country | tuổi thọ của quốc gia tỷ phú |
| tax\_revenue\_country\_  country | doanh thu thuế ở quốc gia tỷ phú |
| total\_tax\_rate\_country | tổng thuế suất tại quốc gia tỷ phú |
| population\_country | dân số của quốc gia tỷ phú |
| latitude\_country | vĩ độ của quốc gia tỷ phú |
| longitude\_country | kinh độ của quốc gia tỷ phú |

* 1. **Chất lượng bộ dữ liệu**
     + - 1. **Phân bố giá trị của các cột có kiểu dữ liệu dạng số**

Với mỗi cột numerical, ta sẽ tính số lượng giá trị thiếu, tỉ lệ % giá trị thiếu (từ 0 đến 100), số lượng giá trị xuất hiện (count), giá trị nhỏ nhất (min), giá trị trung bình (mean), giá trị trung vị (median), giá trị lớn nhất (max), các khoảng tứ phân vị (Q1, Q3) và độ lệch chuẩn (std). Sau đó ta sẽ lưu các giá trị này vào dataframe numerical\_col\_df thành 10 dòng là "num\_missing\_val", "ratio\_missing\_val", "count", "min", "mean", "median", "max", "Q1", "Q3", "std".

Ảnh có chứa văn bản, ảnh chụp màn hình

Mô tả được tạo tự động

Ta rút ra được các kết luận sau:

* Các cột "rank" và "finalWorth" không có xuất hiện các giá trị bị thiếu, trong khi các cột còn lại có xuất hiện giá trị bị thiếu với tỉ lệ từ **6.212%** đến **6.97%** và riêng cột "age" có giá trị bị thiếu với tỉ lệ **2.462%**. Ta sẽ tiền xử lý giá trị bị thiếu của cột này ở các bước sau.
* Cột "rank" có giá trị thấp nhất (min) là **1** và giá trị cao nhất (max) là **2540**, tức là bộ dữ liệu trên chỉ thu thập từ tỷ phú xếp hạng 1 đến hạng 2540 về mức độ giàu có. Ta thấy dữ liệu có tổng cộng **2640** dòng nên điều này có nghĩa là bảng xếp hạng này có thể xét đến trường hợp đồng hạng nhau của các tỷ phú.
* Cột "age" có giá trị trung bình là **65.14**, điều này có nghĩa tuổi thọ trung bình của các tỷ phú là khoảng **65** tuổi. Tỷ phú trẻ tuổi nhất là **18** tuổi (theo giá trị nhỏ nhất) và tỷ phú lớn tuổi nhất là **101** tuổi (theo giá trị lớn nhất).
* Cột "cpi\_change\_country" có chứa giá trị âm **(min = -1.9)**, điều này có nghĩa CPI thay đổi của từng quốc gia có thể tăng hoặc giảm so với năm trước (2022), nếu giá trị âm có nghĩa là đang giảm và ngược lại nếu giá trị dương có nghĩa là đang tăng.
  + - * 1. **Phân bố giá trị của các cột có kiểu dữ liệu dạng phân loại**

Với mỗi cột categorical, ta sẽ tính số lượng giá trị thiếu, tỉ lệ % giá trị thiếu (từ 0 đến 100), số lượng giá trị xuất hiện (count), giá trị xuất hiện nhiều nhất (mode), số lượng giá trị xuất hiện nhiều nhất (không xét giá trị NaN), số lượng các giá trị khác nhau (không xét giá trị NaN) và danh sách các giá trị khác nhau (không xét giá trị NaN). Sau đó ta sẽ lưu các giá trị này vào dataframe categorical\_col\_df thành 7 dòng là "num\_missing\_val", "ratio\_missing\_val", "count", "top", "freq", "nunique", "unique".

Ảnh có chứa văn bản, ảnh chụp màn hình, thực đơn

Mô tả được tạo tự động

Ta rút ra được các kết luận sau:

* Một số cột có giá trị bị thiếu khá ít **(1.44% - 2.879%)** chẳng hạn "country", "city" "birthDate", "birthYear", "birthMonth", "birthDay". Ta để ý thấy số lượng các cột bị thiếu của cột "birthDate" **(76 cột)** bằng với các cột "birthYear", "birthMonth", "birthDay", điều này cũng dễ hiểu.
* Có một số cột có giá trị bị thiếu khá nhiều như "organization", "title", "state", "residenceStateRegion" **(trên 1800 dòng)**. Đối với cột "state" và "residenceStateRegion" chỉ có ý nghĩa khi quốc tịch của các tỷ phú là ở United States. Ta sẽ tìm cách xử lý chúng ở bước tiền xử lý dữ liệu.
* Ở cột "category" và "industries" đều có số lượng giá trị phân biệt (unique) là **18** loại và các giá trị phân biệt này có vẻ giống nhau. Ta có thể loại bỏ đi cột "category" ở bước tiền xử lý dữ liệu.
* Ở cột "date" chỉ có 2 giá trị phân biệt, điều này cho ta biết ngày thu thập dữ liệu chính xác là "2023/04/04" và được thu thập vào lúc "05:01" và "09:01".

# Trực quan hóa dữ liệu

* 1. **Độ tuổi trung bình của các tỷ phú trên thế giới là bao nhiêu?**

Để trả lời cho câu hỏi này, ta sẽ dùng biểu đồ cột để thể hiện số lượng tỷ phú trên thế giới theo từng độ tuổi. Trong đó:

* Trục hoành: thể hiện tuổi (age).
* Trục tung: thể hiện số người theo từng độ tuổi.

Ảnh có chứa Sơ đồ, ảnh chụp màn hình, hàng, biểu đồ

Mô tả được tạo tự động

**Nhận xét**: Quan sát biểu đồ trên, ta nhận thấy:

* Trung bình từ độ tuổi **55 đến 73 tuổi** là độ tuổi có nhiều người tỷ phú nhất. Đây được cho là độ tuổi của những người có nhiều kinh nghiệm nhất.
* Tỷ phú trẻ tuổi nhất là **18 tuổi** và tỷ phú lớn tuổi nhất là **101 tuổi**.

Giả sử nếu ta chia độ tuổi của tỷ phú thành 2 loại: "Tỷ phú dưới 50 tuổi" và "Tỷ phú trên hoặc bằng 50 tuổi". Ta sẽ dùng biểu đồ tròn để thể hiện điều này. Trong đó:

* Màu xanh: thể hiện số lượng tỷ phú có độ tuổi dưới 50.
* Màu cam: thể hiện số lượng tỷ phú có độ tuổi trên hoặc bằng 50.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Phông chữ

Mô tả được tạo tự động

**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy có **89%** những người tỷ phú đều trên hoặc bằng 50 tuổi, còn lại chỉ **11%** những người tỷ phú có độ tuổi nhỏ hơn 50 tuổi.

* 1. **Có nhiều tỷ phú xuất thân tự kinh doanh giữa nam và nữ?**

Để trả lời cho câu hỏi này, ta sẽ dùng 3 loại biểu đồ tròn để thể hiện điều này. Trong đó:

* Màu xanh: thể hiện tỉ lệ tỷ phú có giới tính nam.
* Màu cam: thể hiện tỉ lệ tỷ phú có giới tính nữ.

Mỗi biểu đồ sẽ thể hiện số lượng tỷ phú có tự lập kinh doanh (bên trái) và số lượng tỷ phú không có tự lập kinh doanh (bên phải).

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Phông chữ

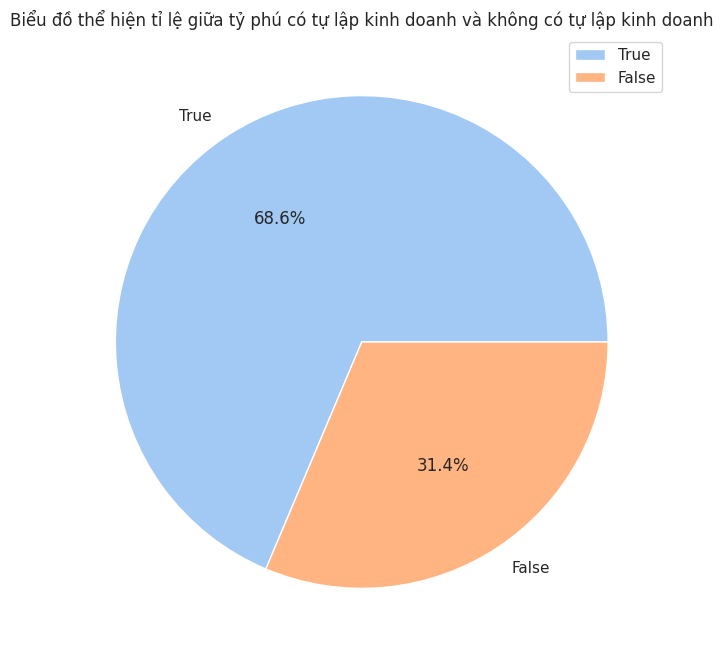
Mô tả được tạo tự động

**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy:

* Đối với những người tỷ phú xuất thân từ việc tự kinh doanh, có thể thấy tỉ lệ nam tỷ phú chiếm áp đảo **(~94.7%)** so với nữ tỷ phú.
* Trong khi đó những người không xuất thân từ việc tự kinh doanh cho thấy nam tỷ phú chỉ chiếm **khoảng 2.4 lần** so với nữ tỷ phú, tuy nhiên trên tỉ lệ giới tính nam là tỷ phú vẫn thống trị phân bố.

Tiếp theo, ta sẽ dùng tiếp 1 biểu đồ tròn để thể hiện tỉ lệ giữa tỷ phú có tự lập kinh doanh và không có tự lập kinh doanh. Trong đó:

* Màu xanh: thể hiện tỉ lệ tỷ phú có tự lập kinh doanh.
* Màu cam: thể hiện tỉ lệ tỷ phú không có tự lập kinh doanh.



**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy đa số khoảng **68.6%** tỷ phú trên thế giới xuất thân từ việc tự kinh doanh.

* 1. **Hiện nay các tỷ phú tập trung nhiều ở các quốc gia nào?**

Để trả lời cho câu hỏi này, ta sẽ dùng biểu đồ map để thể hiện phân bố về các quốc gia mà các tỷ phú đang sinh sống. Trong đó, ta dùng vòng tròn biểu thị màu sắc từ tím tới vàng và kích thước vòng tròn càng lớn đều thể hiện số lượng tỷ phú đang sinh sống ở quốc gia đó theo hướng tăng dần.

Ảnh có chứa bản đồ, văn bản, tập bản đồ

Mô tả được tạo tự động

**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy:

* Mỹ, Trung Quốc, Ấn Độ, Đức, Anh, Nga là các nước có nhiều tỷ phú sinh sống nhất trên thế giới. Trong đó, riêng Mỹ có tới **754** người đang sinh sống và Trung Quốc có tới **523** người đang sinh sống.
* Ngoài ra, có khá nhiều nước ở khu vực châu Phi không có nhiều tỷ phú sinh sống, nếu có thường **không quá 5 người**. Có thể do điều kiện khí hậu ở đây khắc nghiệt hơn chăng.
  1. **Lĩnh vực nào có nhiều nhà tỷ phú đầu tư nhất?**

Để trả lời cho câu hỏi này, ta sẽ dùng biểu đồ cột ngang để thể hiện số lượng tỷ phú trên thế giới theo từng lĩnh vực kinh doanh. Trong đó:

* Trục tung: thể hiện tên lĩnh vực kinh doanh (industries).
* Trục hoành: thể hiện số người tương ứng.

Ảnh có chứa văn bản, hàng, biểu đồ, Sơ đồ

Mô tả được tạo tự động

**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy lĩnh vực được đầu tư nhiều nhất là "Finance & Investments" (Tài chính & Đầu tư) với **372** người, theo sau đó là "Manufacturing" (Sản xuất) và "Technology" (Công nghệ). Có vẻ như ngành "Gambling & Casinos" (Cờ bạc & Sòng bạc) ít được chú ý tới đối với các tỷ phú.

Tương tự, ta cũng sẽ dùng thêm một biểu đồ cột ngang để thể hiện tổng giá trị tài sản ròng của các tỷ phú trên thế giới theo từng lĩnh vực kinh doanh. Trong đó:

* Trục tung: thể hiện tên lĩnh vực kinh doanh (industries).
* Trục hoành: thể hiện tổng giá trị tài sản ròng (finalWorth).

Ảnh có chứa văn bản, hàng, Sơ đồ, biểu đồ

Mô tả được tạo tự động

**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy lĩnh vực có nhiều giá trị tài sản nhất ở đây lại là "Technology" (Công nghệ). Ngược lại, lĩnh vực có ít giá trị tài sản là "Construction & Engineering" (Kiến trúc & Kỹ thuật), theo sau đó là "Gambling & Casinos" (Cờ bạc & Sòng bạc).

* 1. **Có sự khác nhau về lĩnh vực kinh doanh giữa tỷ phú nam và nữ không?**

Để trả lời cho câu hỏi này, ta sẽ dùng 2 biểu đồ cột ngang để thể hiện số lượng tỷ phú trên thế giới (nam và nữ) theo từng lĩnh vực kinh doanh. Trong đó:

* Trục tung: thể hiện tên lĩnh vực kinh doanh (industries).
* Trục hoành: thể hiện số người tương ứng.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, hàng

Mô tả được tạo tự động

**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy:

* Lĩnh vực được nhiều tỷ phú nam đầu tư nhất là "Finance & Investments" (Tài chính & Đầu tư).
* Trong khi đó, lĩnh vực được nhiều tỷ phú nữ đầu tư nhất lại là "Food & Beverage" (Thực phẩm & Nước giải khát). Không có nữ tỷ phú nào trong bảng xếp hạng đầu tư về lĩnh vực "Telecom" (Viễn thông).
  1. **Giá trị tài sản ròng của các tỷ phú được phân bố như thế nào theo quốc tịch?**

Để trả lời cho câu hỏi này, ta sẽ dùng biểu đồ map để thể hiện phân bố về tổng giá trị tài sản ròng của các tỷ phú theo quốc tịch. Trong đó, ta dùng vòng tròn biểu thị màu sắc từ tím tới vàng và kích thước vòng tròn càng lớn đều thể hiện giá trị tài sản ròng tại quốc gia đó càng cao.

Ảnh có chứa bản đồ, tập bản đồ, văn bản

Mô tả được tạo tự động

**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy rằng Mỹ đứng đầu thế giới về tổng giá trị tài sản ròng của các tỷ phú **(4,490,800 triệu đô la Mỹ)**. Tiếp theo là Trung Quốc đứng thứ 2 **(1,648,300 triệu đô la Mỹ)** và Ấn Độ đứng thứ 3 **(674,000 triệu đô la Mỹ)**.

Tiếp theo, ta sẽ dùng biểu đồ histogram để thể hiện phân bố của tài sản ròng của các tỷ phú. Trong đó:

* Trục hoành: thể hiện giá trị tài sản ròng (từ 1000 đến hơn 100000) (finalWorth).
* Trục tung: thể hiện số người tương ứng.

Ảnh có chứa Sơ đồ, biểu đồ, văn bản, hàng

Mô tả được tạo tự động

**Nhận xét:** Quan sát biểu đồ trên, ta nhận thấy khi đồ thị lệch phải nghĩa là giá trị tài sản ròng của các tỷ phú tập trung chủ yếu không trên giá trị trung bình **(mean = 4623.788)**. Giá trị tối thiểu (min) của tài sản ròng đạt **1000** và giá trị tối đa (max) của tài sản ròng đạt **211,000**.

# Rút ra kết luận

1. Độ tuổi trung bình của một tỷ phú là **65 tuổi**. Tỷ phú trẻ tuổi nhất là **18 tuổi** và tỷ phú lớn tuổi nhất là **101 tuổi**.

2. Có đến **89%** những người tỷ phú đều trên hoặc bằng 50 tuổi, còn lại chỉ **11%** những người tỷ phú có độ tuổi nhỏ hơn 50 tuổi.

3. Đa số khoảng **68.6%** tỷ phú trên thế giới xuất thân từ việc tự kinh doanh. Trong đó tỉ lệ nam tỷ phú chiếm áp đảo **(~94.7%)** so với nữ tỷ phú.

4. Mỹ hiện tại là đất nước có nhiều tỷ phú đang sinh sống, sau đó là Trung Quốc, Ấn Độ, Đức, Anh, Nga. Châu Phi không có nhiều điều kiện thích hợp nên có ít tỷ phú sinh sống ở đây.

5. Lĩnh vực được đầu tư nhiều nhất là "Finance & Investments" (Tài chính & Đầu tư) với **372** người và lĩnh vực ít được đầu tư nhất là "Gambling & Casinos" (Cờ bạc & Sòng bạc) đối với các tỷ phú.

6. Tuy nhiên lĩnh vực có nhiều giá trị tài sản nhất ở đây lại là "Technology" (Công nghệ). Ngược lại, lĩnh vực có ít giá trị tài sản là "Construction & Engineering" (Kiến trúc & Kỹ thuật), theo sau đó là "Gambling & Casinos" (Cờ bạc & Sòng bạc).

7. Lĩnh vực được nhiều tỷ phú nam đầu tư nhất là "Finance & Investments" (Tài chính & Đầu tư). Trong khi đó, lĩnh vực được nhiều tỷ phú nữ đầu tư nhất lại là "Food & Beverage" (Thực phẩm & Nước giải khát). Không có nữ tỷ phú nào trong bảng xếp hạng đầu tư về lĩnh vực "Telecom" (Viễn thông).

8. Mỹ đứng đầu thế giới về tổng giá trị tài sản ròng của các tỷ phú **(4,490,800 triệu đô la Mỹ)**. Tiếp theo là Trung Quốc đứng thứ 2 **(1,648,300 triệu đô la Mỹ)** và Ấn Độ đứng thứ 3 **(674,000 triệu đô la Mỹ)**.

9. Tài sản ròng của các tỷ phú trung bình khoảng **4623.788**. Giá trị tối thiểu của tài sản ròng đạt **1000** và giá trị tối đa của tài sản ròng đạt **211,000**.

# Tài liệu tham khảo

* [1] Slide và bài giảng của thầy Nguyễn Tiến Huy.