TRƯỜNG ĐẠI HỌC PHENIKAA

KHOA CÔNG NGHỆ THÔNG TIN

A logo for a university

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN**

**MÔN HỌC: TRỰC QUAN HÓA DỮ LIỆU**

Nhóm 6   
Đề tài: Trực quan hóa dữ liệu Pokemon (Pokemon Index)

Thành viên nhóm

|  |  |  |
| --- | --- | --- |
| Lê Trung Kiên | 21012505 | K15.CNTT5 |
| Đinh Thái Phúc | 21011622 | K15.CNTT |

GVHD: Th.S Nguyễn Văn Thiệu

**05/07/2024 – Hà Nội**

Mục lục

[Danh sách hình vẽ 2](#_Toc171148898)

[Phần 1. Giới thiệu 3](#_Toc171148899)

[**1.1.** **Giới thiệu về trực quan hóa dữ liệu** 3](#_Toc171148900)

[**1.2**  **Vấn đề nghiên cứu** 4](#_Toc171148901)

[**1.3.** **Phương pháp và công cụ sử dụng** 5](#_Toc171148902)

[***1.3.1.*** ***Phương pháp*** 5](#_Toc171148903)

[***1.3.2.*** ***Các thư viện và mô-đun đi kèm*** 5](#_Toc171148904)

[*1.3.3. Jupyter Notebook* 6](#_Toc171148905)

[*1.3.4. Tableau* 6](#_Toc171148906)

[*1.3.5. SQL* 6](#_Toc171148907)

[*1.3.6. Google Colab* 6](#_Toc171148908)

[PHẦN 2. KIẾN THỨC NỀN TẢNG 7](#_Toc171148909)

[2.1. Kiến thức cơ bản về trực quan hóa dữ liệu 7](#_Toc171148910)

[*2.1.1. Khái niệm* 7](#_Toc171148911)

[*2.1.2. Phương pháp* 9](#_Toc171148912)

[*2.1.3. Công cụ quan trọng* 10](#_Toc171148913)

[2.2. Nguyên tắc và quy tắc thiết kế trực quan hóa dữ liệu 11](#_Toc171148914)

[Phần 3. Phân tích dữ liệu dựa vào trực quan hóa 12](#_Toc171148915)

[3.1. Tiền xử lý dữ liệu 12](#_Toc171148916)

[**3.2.** **Phân tích dữ liệu** 14](#_Toc171148917)

[3.3. Trực quan hóa dữ liệu 15](#_Toc171148918)

[*3.3.1. Chỉ số Total* 15](#_Toc171148919)

[*3.3.2. Chỉ số HP* 16](#_Toc171148920)

[*3.3.3. Chỉ số Attack* 17](#_Toc171148921)

[*3.3.4. Chỉ số Defense* 19](#_Toc171148922)

[*3.3.5. Chỉ số Speed* 20](#_Toc171148923)

[*3.3.6. Phân bổ chỉ số và thể hiện sự chênh lệch của các pokemon* 21](#_Toc171148924)

[*3.3.8. Các hệ của Pokemon* 24](#_Toc171148925)

[*3.3.9. Pokemon tốt nhất để đấu với đối thủ (Dữ liệu mở)* 25](#_Toc171148926)

[*3.3.10. Sự phân phối Pokemon của tác giả* 27](#_Toc171148927)

[3.4. POKEDEX 30](#_Toc171148928)

[Phần 4. Kết quả và thảo luận 32](#_Toc171148929)

[**4.1. Các nguyên tắc thiết kế trực quan mà tôi đã áp dụng** 32](#_Toc171148930)

[**4.2. Mô tả kết quả của bài tập lớn** 33](#_Toc171148931)

[**4.3. Thông tin nhận thấy từ trực quan hóa dữ liệu** 34](#_Toc171148932)

[Phần 5. Tổng kết bài tập lớn 35](#_Toc171148933)

[5.1. Kế hoạch 35](#_Toc171148934)

[*5.1.1. Kế hoạch dự kiến* 35](#_Toc171148935)

[*5.1.2. Phân chia công việc* 36](#_Toc171148936)

[5.2. Kết quả đạt được 36](#_Toc171148937)

[5.3. Ưu điểm và hạn chế 37](#_Toc171148938)

[*5.3.1. Ưu điểm* 37](#_Toc171148939)

[*5.3.2. Hạn chế* 37](#_Toc171148940)

[Phần 6. Tài liệu tham khảo 39](#_Toc171148941)

Danh sách hình vẽ

[Hình 1. Các công cụ sử dụng trong python 10](#_Toc171035575)

[Hình 2. Biểu đồ cột thể hiện các giá trị của chỉ số Total. 15](#_Toc171035576)

Hình [3. Biểu đồ cột thể hiện các giá trị của chỉ số HP ... 16](#_Toc171035577)

Hình [4. Biểu đồ cột thể hiện các giá trị của chỉ số Attack 17](#_Toc171035578)

Hình [5. Biểu đồ cột thể hiện các giá trị của chỉ số Defense 19](#_Toc171035579)

Hình [6. Biểu đồ cột thể hiện các giá trị của chỉ số Speed 20](#_Toc171035580)

Hình [7. Biểu đồ phân bổ chỉ số của Pokemon 21](#_Toc171035581)

Hình [8. Biểu đồ sự tương quan chỉ số của Pokemon 23](#_Toc171035582)

Hình [9. Biểu đồ tròn thể hiện các hệ của Pokemon 24](#_Toc171035583)

Hình [10. Biểu đồ Sankey thể hiện các hệ của Pokemon và Pokemon tốt nhất đấu lại chúng 25](#_Toc171035584)

Hình [11. Biểu đồ 3D thể hiện sự phân bố của Pokemon trong thế giới 27](#_Toc171035585)

Hình [12. Biểu đồ KDE thể hiện tỉ lệ xuất hiện của Pokemon huyền thoại 28](#_Toc171035586)

Hình [13. Hình ảnh Pokedex 30](#_Toc171035587)

Danh sách bảng biểu

Bảng 1. Hình ảnh bảng dữ liệu ................................................................................ 14

Bảng 2. Hình ảnh bảng dữ liệu của dữ liệu mở ……………………………………26

Bảng 3. Phân chia công việc ………………………………………………………35

Bảng 4. Bảng nội dung công việc …………………………………………………36

# Phần 1. Giới thiệu

* 1. **Giới thiệu về trực quan hóa dữ liệu**

Trực quan hóa dữ liệu là một lĩnh vực quan trọng trong khoa học dữ liệu, tập trung vào việc sử dụng hình ảnh và đồ họa để biểu diễn dữ liệu phức tạp. Môn học này trang bị cho sinh viên các kỹ năng và kiến thức cần thiết để biến đổi dữ liệu thô thành những biểu đồ, biểu diễn đồ họa trực quan giúp dễ dàng hiểu và phân tích.

Trực quan hóa dữ liệu không chỉ dừng lại ở việc tạo ra những hình ảnh đẹp mắt mà còn đòi hỏi khả năng phân tích, suy luận và kể chuyện thông qua dữ liệu. Sinh viên học môn này sẽ tìm hiểu về các nguyên tắc cơ bản của thiết kế đồ họa, tâm lý học nhận thức, cũng như các công cụ và kỹ thuật hiện đại để tạo ra trực quan hóa dữ liệu hiệu quả.

Vai trò của trực quan hóa dữ liệu trong phân tích dữ liệu và truyền tải thông tin là không thể thay thế. Dưới đây là một số vai trò quan trọng của nó:

* Hỗ trợ ra quyết định: Trực quan hóa dữ liệu giúp các nhà quản lý và nhà nghiên cứu nắm bắt nhanh chóng các xu hướng, mô hình và mối quan hệ trong dữ liệu, từ đó đưa ra các quyết định dựa trên dữ liệu một cách chính xác và hiệu quả hơn.
* Khám phá dữ liệu: Trực quan hóa cho phép các nhà phân tích phát hiện ra các mẫu ẩn giấu, các bất thường và các cơ hội mới trong dữ liệu mà có thể bị bỏ sót nếu chỉ sử dụng các phương pháp phân tích truyền thống.
* Giao tiếp hiệu quả: Các biểu đồ và đồ thị có thể truyền tải thông tin phức tạp một cách trực quan và dễ hiểu hơn so với các bảng số liệu khô khan. Điều này rất quan trọng trong việc trình bày kết quả nghiên cứu cho các đối tượng không chuyên về dữ liệu.
* Tăng cường sự hiểu biết: Trực quan hóa dữ liệu giúp biến những con số khô khan thành những câu chuyện sinh động và có ý nghĩa. Điều này không chỉ giúp người xem hiểu rõ hơn về dữ liệu mà còn tạo ra sự hấp dẫn và thú vị trong việc tiếp cận thông tin.
* Tăng hiệu quả làm việc: Sử dụng các công cụ trực quan hóa dữ liệu giúp tiết kiệm thời gian và công sức trong việc xử lý và phân tích dữ liệu. Các công cụ này thường đi kèm với các chức năng tự động hóa và tương tác, giúp người dùng dễ dàng thực hiện các thao tác phân tích phức tạp.

Trong bối cảnh hiện đại, khi mà lượng dữ liệu ngày càng lớn và phức tạp, trực quan hóa dữ liệu trở thành một kỹ năng không thể thiếu đối với các chuyên gia dữ liệu. Môn học trực quan hóa dữ liệu không chỉ cung cấp kiến thức nền tảng mà còn hướng dẫn sinh viên áp dụng các kỹ thuật này vào các bài toán thực tế, từ đó nâng cao năng lực phân tích và truyền tải thông tin một cách hiệu quả.

**1.2 Vấn đề nghiên cứu**

Báo cáo tập trung và việc phân tích và trực quan hóa dữ liệu Pokemon.

Thế giới Pokemon là một vũ trụ phong phú và đa dạng, được tạo ra bởi Satoshi Tajiri và Ken Sugimori, và ra mắt lần đầu tiên vào năm 1996 bởi công ty Nintendo, Game Freak, và Creatures. Ban đầu, Pokemon bắt đầu như một trò chơi điện tử dành cho hệ máy Game Boy, nhưng sau đó đã phát triển thành một hiện tượng toàn cầu, bao gồm trò chơi điện tử, thẻ bài, loạt phim hoạt hình, phim điện ảnh, truyện tranh, và hàng loạt sản phẩm liên quan.

Pokemon, viết tắt của "Pocket Monsters", là các sinh vật có khả năng đặc biệt và kỹ năng chiến đấu độc đáo. Hiện nay, có hơn 800 loài Pokemon khác nhau, mỗi loài có các thuộc tính, khả năng và đặc điểm riêng biệt. Các Pokemon có thể tiến hóa thành những hình dạng mạnh mẽ hơn và có thể học được nhiều kỹ năng mới qua quá trình huấn luyện.

Pokemon với hệ thống nhân vật phong phú và đa dạng. Mỗi Pokemon có các chỉ số khác nhau như HP, Attack, Defense, và các chỉ số đặc biệt khác. Việc phân tích dữ liệu này sẽ giúp hiểu rõ hơn về đặc điểm và mối quan hệ giữa các chỉ số, cũng như sự phân bố của các loại Pokemon khác nhau.

Thế giới Pokemon là một vũ trụ rộng lớn, phong phú và đầy màu sắc, kết nối mọi người qua nhiều thế hệ và nền văn hóa khác nhau. Với sự phát triển không ngừng và khả năng sáng tạo vô hạn, Pokemon tiếp tục là một hiện tượng toàn cầu, mang lại niềm vui và sự kỳ diệu cho hàng triệu người hâm mộ trên khắp thế giới.

* 1. **Phương pháp và công cụ sử dụng**
     1. ***Phương pháp***

Để thực hiện bài tập lớn này, chúng ta nén url của web([Pokémon Pokédex: list of Pokémon with stats | Pokémon Database (pokemondb.net)](https://pokemondb.net/pokedex/all)) lại bằng cách sử dụng thư viện requests của Python, sau đó lưu trữ nội dung của trang dưới dạng doc. Bằng cách sử dụng doc, chúng ta có thể thu thập tất cả dữ liệu dưới dạng <tr… dưới dạng tr\_elements.

Ngoài ra chúng ta còn sử dụng tài liệu mở để có thể biết được sự phân bố pokemon ở khắp thế giới kì diệu này, để biết nhiều hơn về thế giới của pokemon, chúng ta có thể biết được hệ nào khắc chế hệ nào, nên sử dụng pokemon nào trong giao tranh.

* + 1. ***Các thư viện và mô-đun đi kèm***

Chúng ta sử dụng ngôn ngữ lập trình Python và các thư viện chuyên dụng cho xử lý và trực quan hóa dữ liệu:

- **Pandas:** Dùng để xử lý và phân tích dữ liệu.

- **Matplotlib:** Thư viện cơ bản để tạo các biểu đồ.

- **requests:** sử dụng để thực hiện các yêu cầu HTTP.

- **lxml.html:** làm sạch, thao tác, và trích xuất dữ liệu từ các trang web.

- **itertools:** Mô-đun này chứa các hàm giúp dễ dàng tạo, thao tác và kết hợp các iterators, tiết kiệm bộ nhớ và tăng hiệu suất khi xử lý các dãy dữ liệu lớn.

- **pandas.plotting:** Module này tích hợp chặt chẽ với các cấu trúc dữ liệu của pandas như DataFrame và Series, giúp người dùng dễ dàng tạo ra các biểu đồ trực quan và phân tích dữ liệu một cách hiệu quả

- **pySankey.sankey:** là một thư viện Python được sử dụng để tạo ra các biểu đồ Sankey. Biểu đồ Sankey là một loại biểu đồ dòng chảy được sử dụng để trực quan hóa dòng chảy của một đại lượng (chẳng hạn như năng lượng, tiền, hay tài nguyên) giữa các thực thể khác nhau. Dòng chảy được biểu diễn bằng các đường nối (hoặc dải) có độ rộng tỉ lệ thuận với lượng dòng chảy đó, giúp dễ dàng so sánh và hiểu được các lượng dòng chảy lớn nhỏ giữa các thành phần.

- **sklearn.decomposition:** một mô-đun trong thư viện scikit-learn của Python, được sử dụng để thực hiện các kỹ thuật giảm chiều dữ liệu

- **seaborn:** cung cấp các công cụ để tạo ra các biểu đồ thống kê phức tạp một cách dễ dàng và thẩm mỹ. Nó đặc biệt hữu ích cho việc khám phá và hiểu các mối quan hệ trong dữ liệu

### *1.3.3. Jupyter Notebook*

Cung cấp một môi trường tương tác để viết mã, chạy phân tích và trực quan hóa kết quả. Viết mã Python, thực hiện các phân tích và hiển thị các kết quả trực quan ngay trong cùng một tài liệu.

### *1.3.4. Tableau*

Trực quan hóa dữ liệu phức tạp và tạo các bảng điều khiển tương tác. Tạo các báo cáo trực quan, biểu đồ và bảng điều khiển tương tác để trình bày kết quả phân tích một cách hiệu quả.

### *1.3.5. SQL*

Truy vấn và xử lý dữ liệu trong cơ sở dữ liệu quan hệ. Truy vấn dữ liệu từ các hệ quản trị cơ sở dữ liệu, thực hiện các phép lọc, nhóm và tổng hợp dữ liệu trước khi nhập vào các công cụ phân tích khác.

### *1.3.6. Google Colab*

Colaboratory hay còn gọi là Google Colab, là một sản phẩm từ Google Research, nó cho phép thực thi các câu lệnh Python trên nền tảng đám mây, đặc biệt phù hợp với những bạn nào làm dự án (project) theo nhóm, hoặc muốn chia sẽ file code Python của mình với người khác ví dụ như nhà tuyển dụng. Giao diện của Google Colab rất giống với Jupyter Notebook, một công cụ giúp bạn chạy từng dòng lệnh Python một cách trực quan và kiểm tra kết quả câu lệnh ngay tại chỗ. Cung cấp một môi trường tương tác để viết mã, chạy phân tích và trực quan hóa kết quả. Viết mã Python, thực hiện các phân tích và hiển thị các kết quả trực quan ngay trong cùng một tài liệu.

# PHẦN 2. KIẾN THỨC NỀN TẢNG

## 2.1. Kiến thức cơ bản về trực quan hóa dữ liệu

### *2.1.1. Khái niệm*

Trực quan hóa dữ liệu là quá trình sử dụng các yếu tố hình ảnh như đồ thị, biểu đồ hoặc bản đồ để trình bày dữ liệu. Quá trình này chuyển đổi dữ liệu phức tạp, có dung lượng lớn hoặc dữ liệu số thành hình ảnh trình bày trực quan có thể xử lý dễ dàng hơn. Các công cụ trực quan hóa dữ liệu cải thiện và tự động hóa quá trình giao tiếp bằng hình ảnh nhằm đảm bảo độ chính xác và chi tiết. Bạn có thể sử dụng những hình ảnh trình bày trực quan để trích xuất những thông tin chuyên sâu hữu ích từ dữ liệu thô.

*Các thành phần của trực quan hóa dữ liệu:*

- Thông điệp: Muốn trình bày mục đích gì?

**VD**: Đo lường hiệu suất công việc hoặc doanh số bán hàng.

- Dữ liệu: sau khi có thông điệp, cần thu thập dữ liệu thích hợp. Sau đó, chỉnh sửa các định dạng dữ liệu hiện có, làm sạch dữ liệu, loại bỏ các giá trị ngoại lai và thực hiện phân tích sâu hơn. Sau đó, lên kế hoạch sử dụng các phương pháp khác nhau để khám phá trực quan.

- Phương tiện trực quan: Lựa chọn công cụ sau đó tạo biểu đồ làm nổi bật các điểm chính. Đơn giản hóa tập dữ liệu phức tạp và trình bày có hệ thống.

*Lợi ích của trực quan hóa dữ liệu:*

- Đưa ra quyết định chiến lược: Người có quyền đưa ra quyết định tiết kiệm thời gian, phân tích, đánh giá tổng quát hóa toàn cảnh.

- Cải thiện dịch vụ khách hàng: Làm nổi bật nhu cầu và mong muốn của khách hàng thông qua biểu diễn đồ họa, xác định những lỗ hổng trong dịch vụ khách hàng, cải thiện sản phẩm hoặc dịch vụ theo chiến lược và giảm hoạt động kém hiệu quả.

- Tăng mức độ tương tác của nhân viên: Tăng khả năng truyền đạt, dễ dàng biểu diễn kế hoạch và mục tiêu chung. Dễ dàng đo lường mục tiêu, tiến độ cũng như sự cải thiện của nhóm.

*Các bước trực quan hóa dữ liệu:*

- Xác định mục tiêu: Xác định các câu hỏi mà tập dữ liệu có thể trả lời.

- Thu thập dữ liệu: xác định nguồn dữ liệu bên trong và bên ngoài. Có sẵn các tập dữ liệu lớn trực tuyến để mua và sử dụng. Hoặc công ty cũng có thể có sẵn các kho lưu trữ dữ liệu hiện có để phân tích.

- Làm sạch dữ liệu: Làm sạch dữ liệu liên quan đến việc loại bỏ dữ liệu dư thừa. Thực hiện các phép tính toán để phân tích thêm hoặc lọc và chuyển đổi dữ liệu để đáp ứng các tiêu chí của câu hỏi.

- Chọn phương tiện trực quan hóa dữ liệu: Tùy thuộc vào mối quan hệ giữa dữ liệu và thông tin muốn truyền tải. Có thể chọn các loại biểu đồ khác nhau.

+ Trực quan hóa tĩnh

+ Trực quan hóa tương tác

- Tạo phương tiện trực quan hóa dữ liệu: sử dụng các công cụ trực quan hóa dữ liệu.

### *2.1.2. Phương pháp*

*Có 5 phương pháp chính, phụ thuộc vào mục đích:*

- Comparison: nếu chúng ta muốn biết sự hơn kém giữa các đối tượng nghiên cứu.

- Composition: nếu chúng ta muốn biết cụ thể về từng thành phần cấu tạo nên một thứ gì đó, một nhóm, v.v... và mỗi thành phần ấy chiếm tỷ lệ như thế nào.

- Distribution: nếu chúng ta muốn xem xét về phân phối dữ liệu, mức độ phân tán, mức độ tập trung của dữ liệu, hãy xác định các điểm bất thường, giá trị ngoại lệ của dữ liệu.

- Trend: nếu chúng ta muốn tìm hiểu về xu hướng biến động của đối tượng nghiên cứu trong dữ liệu.

- Relationship/Correlation: nếu chúng ta muốn tìm hiểu về mối liên hệ, tương quan giữa 2 hay nhiều đối tượng nghiên cứu trong dữ liệu

*Các biểu đồ thường được sử dụng trong trực quan hóa dữ liệu:*

- Biểu đồ cột, đường, hình tròn, … để thể hiện xu hướng, so sánh giá trị.

- Đồ thị để thể hiện mối quan hệ giữa các yếu tố dữ liệu.

- Bản đồ heatmap để thể hiện mức độ tập trung của dữ liệu

### *2.1.3. Công cụ quan trọng*

**A diagram of a programming language

Description automatically generated**

Hình 1. Các công cụ sử dụng trong python

- Google data studio / Looker studio: được sử dụng để có thể kết nối dễ dàng nhiều sản phẩm, sử dụng miễn phí.

- Tableau: giúp người dùng có thể dễ dàng sắp xếp dữ liệu và đầu vào từ nhiều nguồn đồng thời có thể tích hợp nhiều nền tảng khác.

- Infogram: có nhiều tùy chọn độc đáo, nhiều mẫu tối ưu hóa cho di động và mạng xã hội. Hoàn toàn có thể nhúng được vào các thiết kế trang web.

- D3.js: được sử dụng để tích hợp tính năng như mã nguồn mở, nhiều ví dụ tương tác, phân cấp và hoạt ảnh mạnh mẽ.

- Excel, PowerBI là bộ công cụ chuyên về DV của Microsoft, được sử dụng phù hợp các phân tích ngắn.

- Ngôn ngữ lập trình Python và các thư viện liên quan:

* Miễn phí, dễ học, dễ thực hiện, nhiều thư viện hỗ trợ.
* Có rất nhiều thư viện để người dùng có thể sử dụng như Matplotlib, Seaborn, Pandas, GeoPandas,…

## 2.2. Nguyên tắc và quy tắc thiết kế trực quan hóa dữ liệu

Các nguyên tắc và quy tắc thiết kế để trực quan hóa dữ liệu hiệu quả:

- Thu hút khán giả chú ý đến các chi tiết quan trọng thông qua kích cỡ, màu sắc, phông chữ và đồ họa.

- Cung cấp ngữ cảnh cho dữ liệu bằng các dấu hiệu trực quan.

- Chọn kiểu phối màu phù hợp cho phân tích dữ liệu.

- Sử dụng tiêu đề giải thích để cung cấp thông tin chuyên sâu chính cho khán giả và giúp họ tập trung vào đúng câu hỏi.

- Thêm nhãn và số rõ ràng cho dữ liệu.

- Chọn đúng loại biểu đồ: “Chức năng đi trước, trình bày đi sau”. Mục đích là cần được cân nhắc đầu tiên. Đặt câu hỏi để chọn biểu đồ phù hợp: Mục đích là gì? So sánh giá trị gì? Thể hiện xu hướng hay tìm hiểu phân phối và mối quan hệ giữa các biến.

- Không phải tất cả các data đều quan trọng: Không nên ném toàn bộ dataset vào 1 biểu đồ. Phân tích trước để biểu diễn những thông tin quan trọng cho người xem dễ nắm bắt.

- Biểu đồ thể hiện đúng tương quan số liệu thực tế: Không nên lợi dụng các yếu tố về khung tham chiếu, tỉ lệ hình ảnh để gây nhầm lẫn cho người xem.

Đưa ra những kiến thức cơ bản về trực quan hóa dữ liệu, bao gồm các khái niệm, phương pháp và công cụ quan trọng.

Trình bày các nguyên tắc và quy tắc thiết kế trực quan hóa dữ liệu hiệu quả.

# Phần 3. Phân tích dữ liệu dựa vào trực quan hóa

## 3.1. Tiền xử lý dữ liệu

Trong bài tập lớn về trực quan hóa dữ liệu, nhóm chúng em đã sử dụng kiến thức cơ bản về Web Scraping (Quét Web), DataFrames (Khung dữ liệu), và Plotting (Vẽ đồ thị) để thu thập, xử lý và phân tích dữ liệu Pokémon. Quá trình này bao gồm thu thập dữ liệu từ cơ sở dữ liệu trực tuyến, làm sạch dữ liệu để loại bỏ các yếu tố không cần thiết, và chuyển đổi dữ liệu thành định dạng phù hợp để dễ dàng phân tích.

- Thu thập dữ liệu:

Đầu tiên, nhóm đã xác định URL của trang web [Pokemon Database] (<https://pokemondb.net/pokedex/all>) để thu thập dữ liệu về các Pokémon. Đây là trang cung cấp thông tin chi tiết về từng Pokémon, bao gồm tên, hệ, chỉ số, và các thuộc tính quan trọng khác.

+ Khám phá cấu trúc trang web:

Nhóm đã sử dụng công cụ 'Inspect element' để kiểm tra cấu trúc HTML của trang web. Qua việc kiểm tra này, chúng em nhận thấy rằng dữ liệu về các Pokémon được lưu trữ trong các thẻ `<tr>` của bảng HTML, với mỗi hàng `<tr>` đại diện cho một Pokémon. Các thẻ `<td>` bên trong `<tr>` chứa các thông tin chi tiết.

+ Thu thập nội dung trang web:

Để lấy nội dung của trang, nhóm đã sử dụng thư viện `requests` để truy cập và tải nội dung của trang web về. Nội dung này sau đó được chuyển đổi thành dạng cây HTML để dễ dàng trích xuất các thẻ `<tr>` chứa dữ liệu

- Làm sạch dữ liệu:

+ Kiểm tra kích thước hàng:\*\* Sau khi thu thập được các phần tử `<tr>`, nhóm đã kiểm tra kích thước của 12 hàng đầu tiên để đảm bảo rằng các hàng đều có độ dài nhất quán, phù hợp với cấu trúc của bảng dữ liệu. Việc này giúp xác nhận rằng dữ liệu thu thập được là từ bảng chứa thông tin về Pokémon và không bị trộn lẫn với dữ liệu khác trên trang.

+ Trích xuất dữ liệu:

Dựa trên cấu trúc HTML đã được xác định, nhóm tiến hành trích xuất từng cột dữ liệu từ các thẻ `<tr>`. Mỗi thẻ `<td>` bên trong `<tr>` chứa một mục dữ liệu cụ thể (chẳng hạn như tên Pokémon, hệ, chỉ số), và dữ liệu này được thêm vào các danh sách tương ứng trong bộ dữ liệu.

+ Làm sạch các ký tự không mong muốn:

Trong quá trình thu thập dữ liệu, nhóm nhận thấy có sự xuất hiện của các ký tự không mong muốn như '\n' trong dữ liệu. Để làm sạch dữ liệu, chúng em đã loại bỏ các ký tự này nhằm đảm bảo rằng dữ liệu được định dạng chính xác và dễ dàng xử lý trong các bước tiếp theo.

- Chuyển đổi dữ liệu:

+ Tạo DataFrame: Sau khi dữ liệu đã được làm sạch, nhóm đã chuyển đổi dữ liệu thành một DataFrame của Pandas. Đây là một cấu trúc dữ liệu mạnh mẽ trong Python, giúp dễ dàng xử lý và phân tích dữ liệu. DataFrame chứa các cột tương ứng với các thuộc tính của Pokémon và các hàng tương ứng với từng Pokémon.

+ Chuyển đổi kiểu dữ liệu:

Một số cột dữ liệu chứa các chỉ số của Pokémon cần được chuyển đổi từ định dạng chuỗi sang định dạng số. Điều này giúp thực hiện các phép toán và phân tích một cách chính xác, chẳng hạn như tính toán trung bình hoặc so sánh các chỉ số.

+ Lưu trữ dữ liệu:

Cuối cùng, DataFrame này được lưu trữ dưới định dạng file CSV. Điều này giúp nhóm dễ dàng truy cập và sử dụng dữ liệu cho các bước phân tích và trực quan hóa tiếp theo, đồng thời đảm bảo dữ liệu được lưu trữ một cách an toàn và có thể tái sử dụng.

* Trình bày các phân tích và khám phá dữ liệu mà bạn đã thực hiện trong bài tập lớn.
* Mô tả quá trình thiết kế các biểu đồ, đồ thị hoặc trực quan hóa dữ liệu khác. Vẽ biểu đồ này thể hiện điều gì, đưa ra đánh giá nhận xét về biểu đồ, nó cho ta thấy điều gì?
* Trình bày các nguyên tắc thiết kế trực quan mà bạn đã áp dụng và lý do chọn lựa.
  1. **Phân tích dữ liệu**

A screenshot of a computer

Description automatically generated

Bảng 1. Hình ảnh bảng dữ liệu

**Cấu trúc Dữ liệu:**

* Name: Tên của pokemon
* Type: Hệ của pokemon
* Total: Tổng giá trị chỉ số của Pokemon
* Hp: Chỉ số máu của pokemon
* Attack: Chỉ số tấn công của pokemon
* Defense: Chỉ số phòng thủ của pokemon
* Sp. Atk: Chỉ số tấn công đặc biệt của pokemon khi sử dụng kĩ năng đặc biệt
* Sp. Def: Chỉ số phòng thủ đặc biệt của pokemon chống đỡ lại kĩ năng đặc biệt của đối thủ
* Speed: Chỉ số tốc độ

**Nhận xét về dữ liệu:**

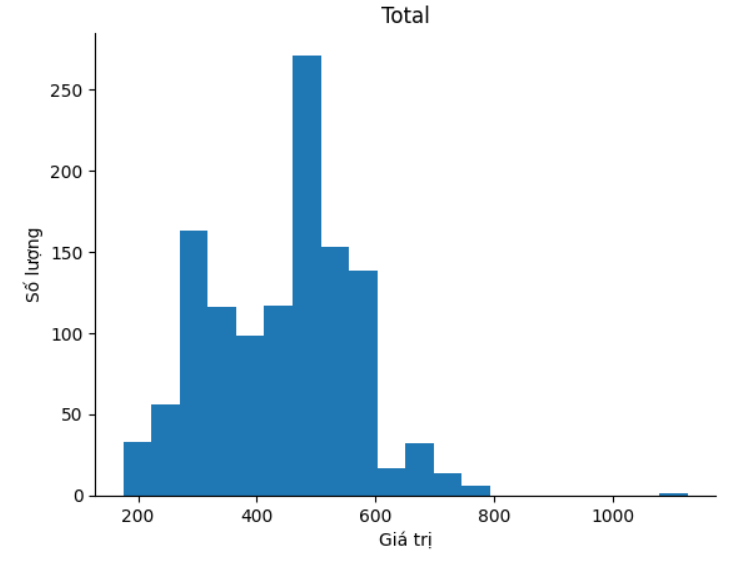
* Chỉ số HP: Chỉ số có sự biến động ít nhất, phản ánh chỉ số máu của tất cả pokemon là same nhau
* Chỉ số Attack, Sp. Attack, Speed: Các chỉ số có sự biến động lớn nhất
* Chỉ số Defense, Sp. Defense: 2 chỉ số này ít biến động nhưng vẫn có những phạm vi tương đối rộng

**Tổng Kết:**

Dữ liệu cho ta thấy mỗi Pokemon mang tộc hệ riêng và các chỉ số của bản thân chúng là khác nhau, có những con rất mạnh về tấn công nhưng lại yếu về phòng thủ. Có những sinh vật mạnh cả 2 thì thường là hiếm hoặc bản thân chúng sinh ra đã được mặc định là sẽ hoàn hảo như vậy.

## 3.3. Trực quan hóa dữ liệu

### *3.3.1. Chỉ số Total*



Hình 2. Biều đồ cột thể hiện các giá trị của chỉ số Total

***Trục hoành (x-axis):***

Trục này biểu diễn giá trị của chỉ số Total của tất cả Pokemon

***Trục tung (y-axis):***

Trục này biểu diễn số lượng pokemon đạt đến ngưỡng giá trị

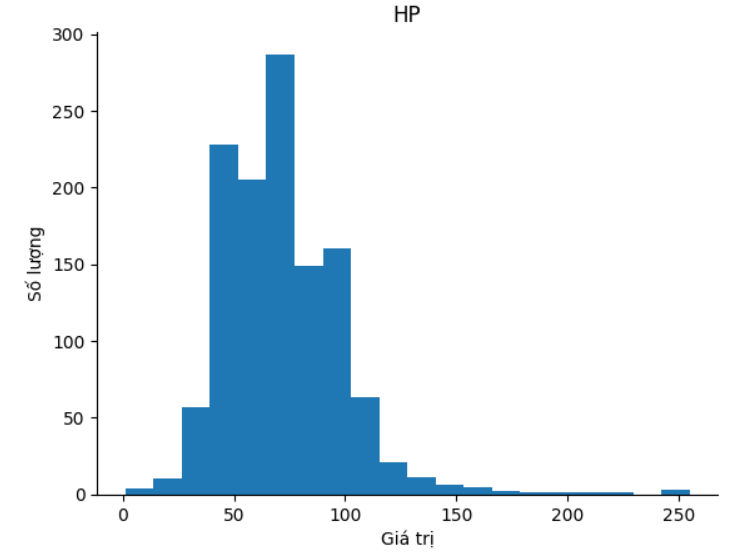
***Nhận xét các giá trị :***

- Bảng này cho ta thấy số lượng pokemon có chỉ số total từ 400 đến 600 chiếm đa số

- Giữa khoảng 800 - 1000 dường như ko có pokemon nào đạt mức này

- Và lớn hơn 1000 chỉ có 1 pokemon đạt đến ngưỡng này.

### *3.3.2. Chỉ số HP*



Hình 3. Biều đồ cột thể hiện các giá trị của chỉ số HP

***Trục hoành (x-axis):***

Trục này biểu diễn giá trị của chỉ số HP của tất cả Pokemon

***Trục tung (y-axis):***

Trục này biểu diễn số lượng pokemon đạt đến ngưỡng giá trị

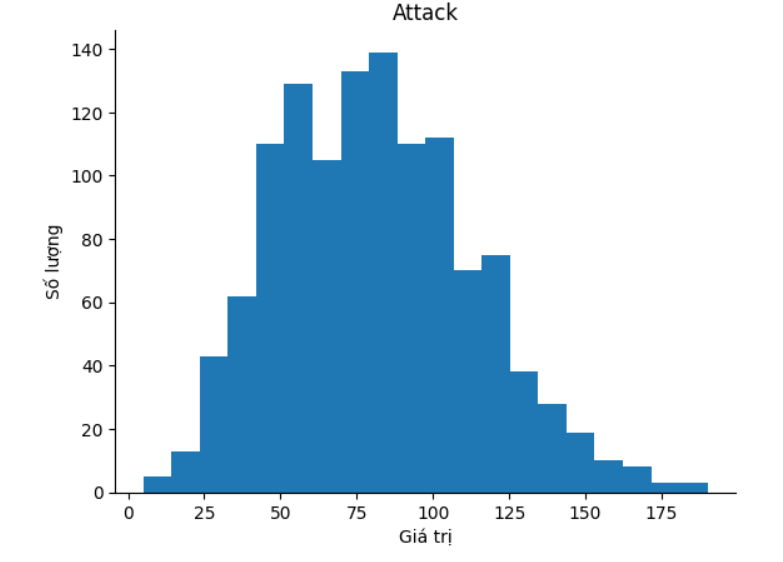
***Nhận xét các giá trị :***

- Số lượng pokemon có chỉ số HP từ 50 đến 100 chiếm đa số

- Lớn hơn 255 không có pokemon nào

- Từ khoảng 150 đến ngưỡng 250 (và lớn hơn 1 chút) chỉ có khoảng 20 con pokemon đạt được đến mức độ chỉ số này

### *3.3.3. Chỉ số Attack*



Hình 4. Biều đồ cột thể hiện các giá trị của chỉ số Attack

***Trục hoành (x-axis):***

Trục này biểu diễn giá trị của chỉ số Attack của tất cả Pokemon

***Trục tung (y-axis):***

Trục này biểu diễn số lượng pokemon đạt đến ngưỡng giá trị

***Nhận xét các giá trị :***

- Số lượng pokemon có chỉ số Attack từ 50 đến 100 chiếm đa số

- Pokemon có chỉ số attack lớn hơn 190 và nhỏ hơn 5 là không tồn tại

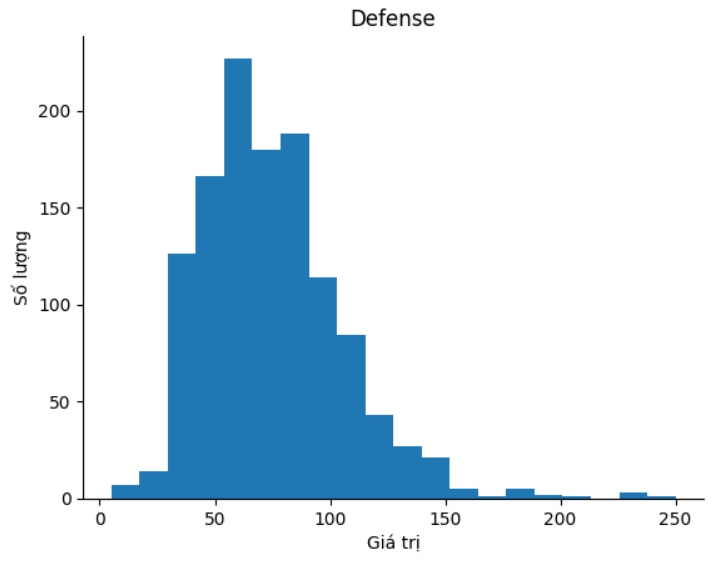
- Pokemon có chỉ số attack nhỏ hơn 25 chiếm thiểu số

- Bắt đầu từ 125 đổ đi ta thấy số lượng pokemon có atk mạnh bắt đầu giảm dần

- Nhưng biểu đồ cho ta thấy số lượng pokemon có chỉ số atk khỏe không quá hiếm như 2 chỉ số ở trên là HP và Total

- Không hiếm nhưng cũng không bị tràn lan quá, điều này tạo 1 môi trường Pokemon có sự phân chia rõ ràng hơn

### *3.3.4. Chỉ số Defense*



Hình 5. Biều đồ cột thể hiện các giá trị của chỉ số Defense

***Trục hoành (x-axis):***

Trục này biểu diễn giá trị của chỉ số Total của tất cả Pokemon

***Trục tung (y-axis):***

Trục này biểu diễn số lượng pokemon đạt đến ngưỡng giá trị

***Nhận xét các giá trị :***

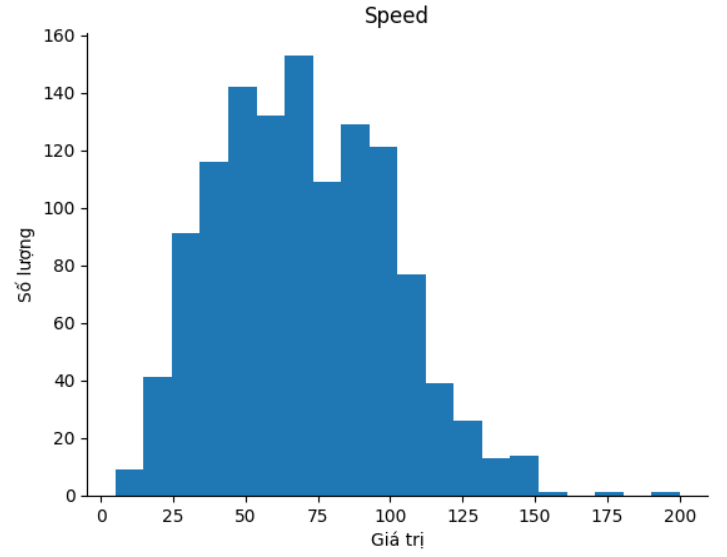
- Số lượng pokemon có chỉ số Defense từ 50 đến 100 chiếm đa số

- Từ khoảng chỉ số 100 đổ đi ta thấy số lượng đạt mức độ phòng thủ này bắt đầu giảm đi

- từ khoảng 150 đến 250 thì số lượng rải rác và không còn phổ biến nữa

- Và Defense điểm 250 là điểm cao nhất của 1 pokemon có thể có

### *3.3.5. Chỉ số Speed*



Hình 6. Biều đồ cột thể hiện các giá trị của chỉ số Speed

***Trục hoành (x-axis):***

Trục này biểu diễn giá trị của chỉ số Total của tất cả Pokemon

***Trục tung (y-axis):***

Trục này biểu diễn số lượng pokemon đạt đến ngưỡng giá trị

***Nhận xét các giá trị :***

- Chỉ số speed là 1 trong những yếu tố quan trọng của 1 ván đầu

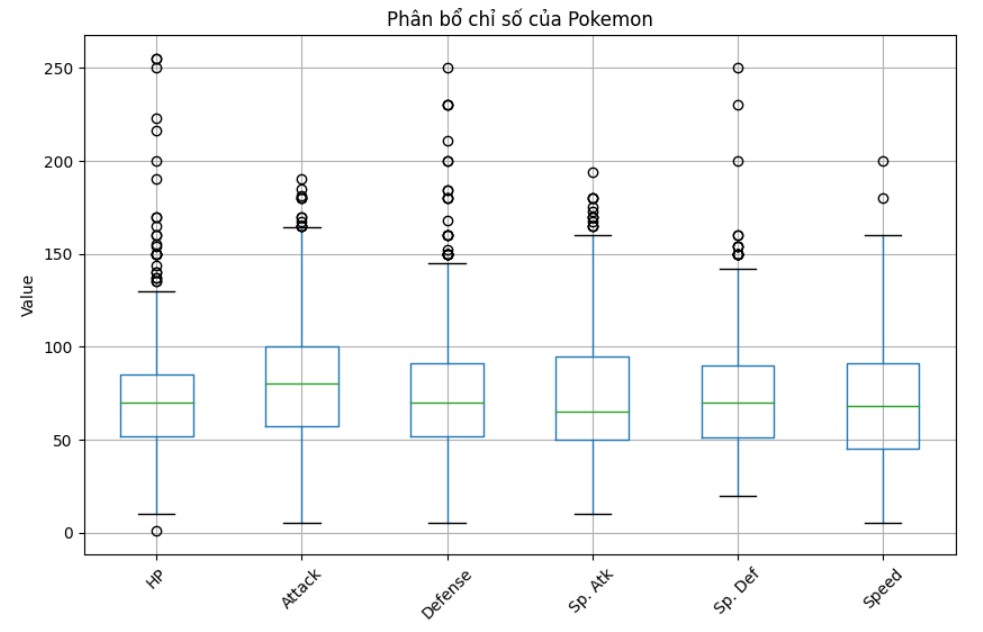
- Tuy pokemon của bạn có atk thấp hay chỉ số def thấp, cũng có thể tạo ra 1 số tình huống lật kèo đáng nhớ dành cho các pokemon thủ

- Chỉ số spd nhỏ hơn 150 không hiếm

- chiếm đa số,phổ biến và hay xuất hiện sẽ là từ khoảng 25 đến 100

- Không quá 10 con pokemon có chỉ số spd lớn lớn 150

### *3.3.6. Phân bổ chỉ số và thể hiện sự chênh lệch của các pokemon*



Hình 7. Biều đồ phân bổ chỉ số của Pokemon

**Sự tương đồng :**

Nhìn chung, ta có thể thấy tất cả các giá trị trung bình của tất cả chỉ số đều nằm trong khoảng từ 50 đến 100, riêng chỉ số Speed thì những pokemon có giá trị trung bình sẽ <50 chiếm 1 chút ít.

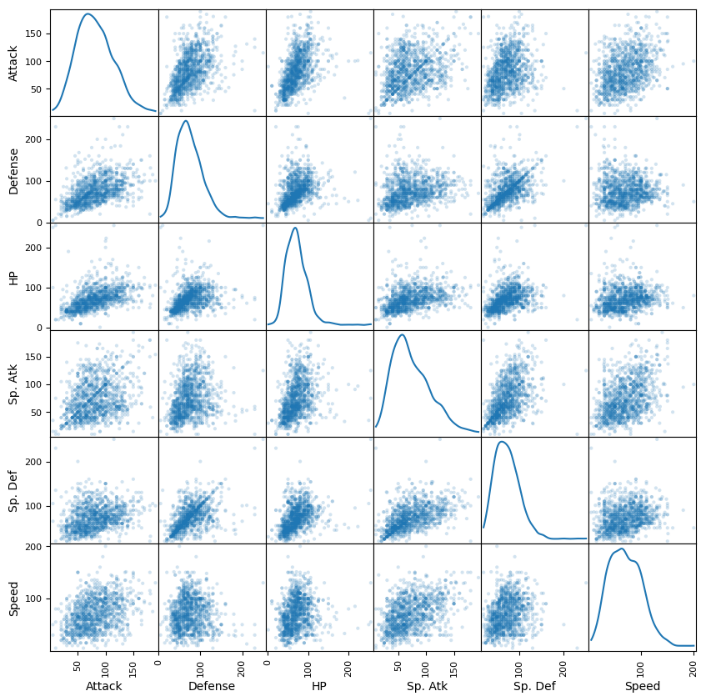
**Sự chênh lệch trong nội bộ từng chỉ số:**

* Chỉ số HP: Chỉ số có sự biến động ít nhất, phản ánh chỉ số máu của tất cả pokemon là same nhau
* Chỉ số Attack, Sp. Attack, Speed: Các chỉ số có sự biến động lớn nhất
* Chỉ số Defense, Sp. Defense: 2 chỉ số này ít biến động nhưng vẫn có những phạm vi tương đối rộng

**Các giá trị ngoại lệ**

* Chỉ số HP, Defense, Sp. Defense: Các chỉ số này có nhiều giá trị ngoại lệ cao, thường thì những pokemon hệ hiếm mới đạt được ngường này

- Chỉ số Attack, Sp. Attack, Speed: Ít giá trị ngoại lệ hơn, giá trị chủ yếu tập trung trong một khoảng hẹp

***3.3.7. Sự tương quan của các chỉ số***

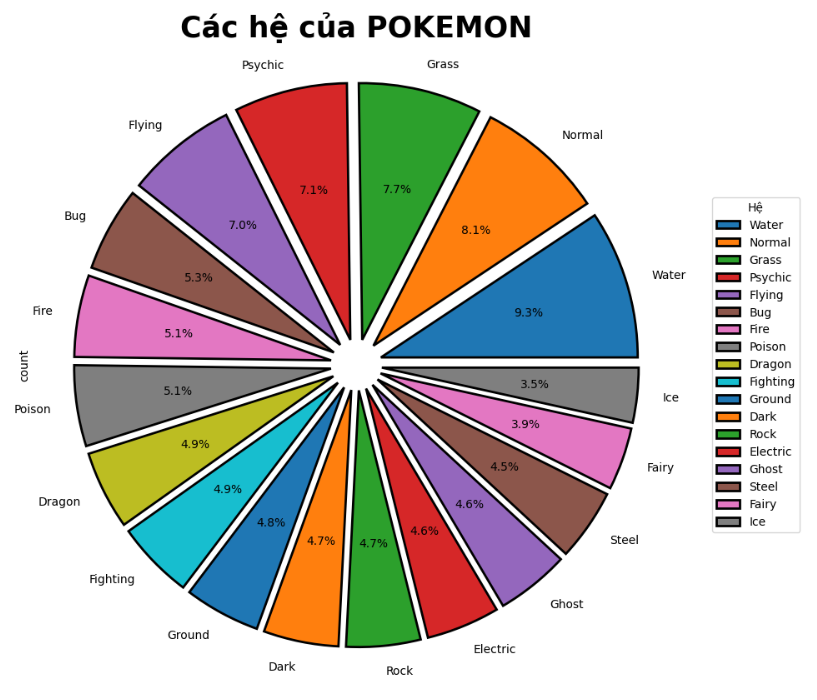
Hình 8. Biều đồ sự tương quan chỉ số của Pokemon

* Ta có thể thấy Defense và Special Defense là 1 mối quan hệ tương quan đồng biến, cả 2 đều phát triển theo cùng 1 hướng
* Các số liệu thống kê khác cũng có một số mối liên kết nhưng giá trị R^2 quá thấp.

**Kết luận :**

Biểu đồ cho thấy một bức tranh tổng thể về sự tiến bộ và phát triển của Tour de France qua các năm, với xu hướng tăng dần trong tốc độ trung bình của người chiến thắng. Điều này phản ánh sự cải tiến liên tục trong công nghệ, phương pháp huấn luyện và các yếu tố khác ảnh hưởng đến khả năng thi đấu của các vận động viên.

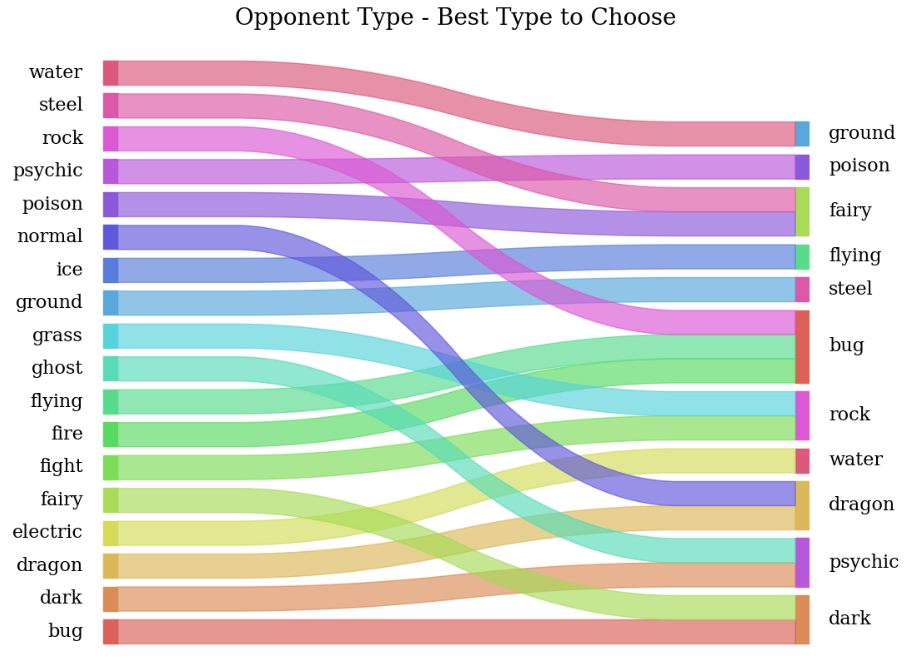
### *3.3.8. Các hệ của Pokemon*



Hình 9. Biều đồ tròn thể hiện các hệ của Pokemon

* Ta có thể nhận thấy, có 1215 pokemon khác nhau với 18 tộc hệ riêng biệt
* Số lượng Pokemon hệ nước(water) chiếm nhiều nhất là 9.3%
* Số lượng Pokemon hệ băng(ice) chiếm thiểu số chỉ vỏn vẹn 3.5% trên tổng số
* Psychic và Flying có thể coi là gần bằng nhau khi chỉ cách nhau 0.1%
* Bug, Fire, Poison là 3 hệ có đầu là 5
* Dragon, Fighting, Ground, Dark, Rock, Electric, Ghost, Steel là các hệ phổ biến vì có đầu 4 là nhiều, có vẻ những hệ này không quá khó và cũng không quá dễ để sở hữu, bạn sẽ có 4.5% pokemon để sở hữu hệ steel và 4.9% đối với pokemon hệ rồng(dragon)

### *3.3.9. Pokemon tốt nhất để đấu với đối thủ (Dữ liệu mở)*



Hình 10. Biều đồ Sankey thể hiện các hệ của Pokemon và Pokemon tốt nhất đấu lại chúng

* Đây là 1 trong những biểu đồ đẹp nhất, bạn có thể nhận thấy, nhìn chúng giống như 1 cầu vồng có nhiều hơn 7 sắc vậy.
* Bên trái là hệ của pokemon đối thủ, và bên phải là pokemon hệ tốt nhất để đem ra đấu với chúng
* Ta có thể dễ dàng nhận thấy, pokemon hệ "bug" là hữu dụng nhất khi có thể đem ra sử dụng vì nó có thể đấu tốt nhất với 3 hệ là flying, fire, rock.

- Trong quá trình hình thành nên biểu đồ này, chúng ta cần load data trước và xem bên trong nó có những biến gì, sau khi mổ tung ra và chúng ta chỉ cần quan tâm đến các biến giúp ích cho quá trình tạo nên biểu đồ liên quan đến vấn đề ta quan tâm.

- Ta chỉ cần quan tâm đến “types” và các biến hệ của pokemon rồi chúng ta sẽ tổng hợp lại thành 1 dataframe mới rồi cho chúng đối đầu với nhau

- Sau đấy ta nhóm khung dữ liệu theo 'type1' và điểm trung bình của từng loại so với against\_x ( x ở đây là hệ của chúng)

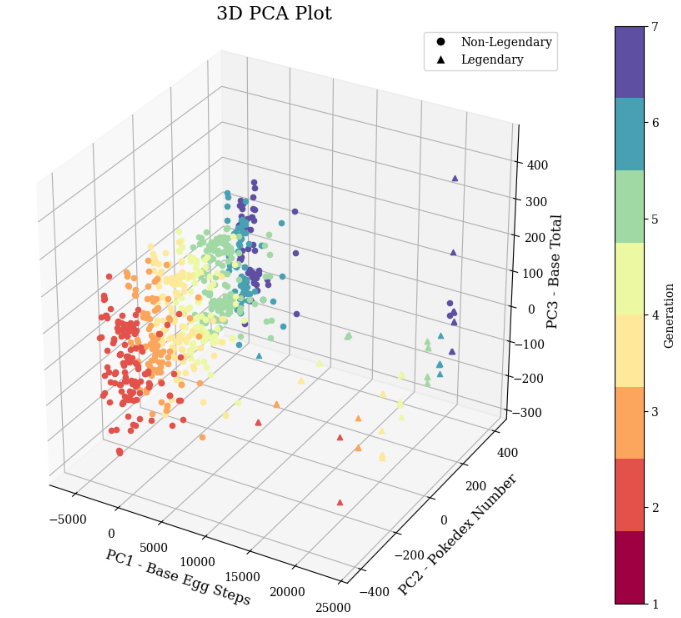
A screenshot of a computer

Description automatically generated

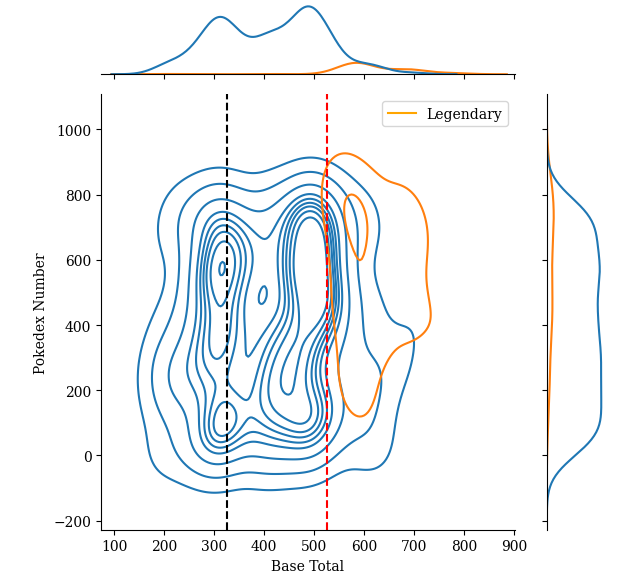
Bảng 2. Hình ảnh bảng dữ liệu của dữ liệu mở

* **Type1**: loại của pokemon
* **against\_x**: x ở là hệ của pokemon

### *3.3.10. Sự phân phối Pokemon của tác giả*



Hình 11. Biều đồ 3D thể hiện sự phân bố của Pokemon trong thế giới

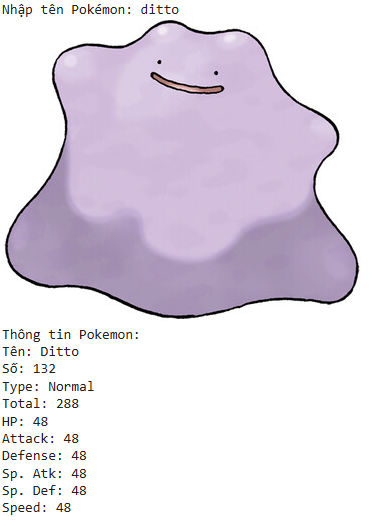


Hình 12. Biều đồ KDE thể hiện tỉ lệ xuất hiện của Pokemon huyền thoại

***Giải thích cho hình 11-12:***

* Giả sử bạn không có thời gian để xem qua tất cả Pokemon của mình, người dùng sẽ muốn có hướng dẫn chẩn đoán nhanh để xem Loại Pokemon nào sẽ là tốt nhất để chọn trong bất kỳ trường hợp cụ thể nào để đấu với một loại Pokemon cụ thể.
* Để thể hiện rõ nhất điều này, chúng tôi đã tạo một khung dữ liệu mới chỉ xem xét các cột 'type1' và 'against\_x'.
* Điều này dẫn đến Sơ đồ Sankey dễ đọc để cung cấp cho chúng tôi loại pokemon tốt nhất để đánh bại đối thủ
* Rõ ràng là có sự phân bổ hai chiều trong Base Total.
* Pokemon huyền thoại được xác định bởi số lượng Base Egg Steps và Base Total.
* Hơn nữa, màu sắc của thế hệ cho thấy pokemon huyền thoại KHÔNG tập trung ở bất kỳ thế hệ cụ thể nào mà lan rộng khắp Pokedex.
* PCA và KDE cho thấy rõ ràng rằng Pokemon huyền thoại nằm trong một vùng của riêng chúng. Tất cả chúng không chỉ tập trung vào phần trên trung bình của KDE, mà còn chủ yếu bao gồm phần lớn Pokemon có 20.000 bước trở lên và có base total cao. Thang đo Pokedex cho thấy có sự phân bố Pokemon Huyền thoại tương đối giống nhau trên mỗi thế hệ, phân bố đều trên các số Pokedex.
* Có khoảng 50-50 cơ hội nhận được pokemon trên trung bình hoặc dưới trung bình, được ước tính bởi một khu vực tương tự dưới đường cong KDE.
* Với base total trên mức trung bình, những Pokemon huyền thoại này rất mạnh và có thể đánh bại hầu hết các đối thủ chỉ dựa vào chỉ số.
* Khi xem Type và Type Sankey của chúng tôi, chúng ta có thể biết loại Pokemon nào có khả năng đánh bại loại đối thủ mạnh nhất. Do đó, tất cả các chỉ số khác đều như nhau, một số đòn tấn công có thể hiệu quả hơn những đòn tấn công khác tùy thuộc vào loại Pokemon. Dữ liệu bổ sung này có thể giúp người dùng đưa ra quyết định sáng suốt về việc nên chọn Pokemon nào trong danh sách của họ

## 3.4. POKEDEX

****

Hình 13. Hình ảnh Pokedex

- Một cái tên quá quen thuộc với những fan của tựa phim pokemon, ở trong thế giới pokemon. Chúng ta có thể tra thông tin của những con poke vừa được thu phục hoặc những con ta chưa sở hữu thông qua 1 máy đỏ kiểu dáng điện thoại màu đỏ.

- Với Pokedex chúng ta có thể dễ dàng tra thông tin, chỉ số của những poke ta đang lưu ý đến.

- Google Colab không hỗ trợ trực tiếp việc tạo GUI truyền thống như có thể làm với Tkinter trong Python. Thế nên em sử dụng các thư viện như "ipywidgets" và "IPython.display" để tạo các giao diện tương tác cơ bản trong môi trường Jupyter Notebook, bao gồm cả Google Colab.

+ ipywidgets : giúp tạo các widget tương tác trong các ô notebook. Những widget này cung cấp giao diện người dùng đồ họa (GUI) để tương tác với mã Python và dữ liệu trực tiếp trong notebook.

+ IPython.display : cung cấp các công cụ để hiển thị và điều khiển các kiểu dữ liệu khác nhau trong notebook.

Như đã thấy, với Pokedex ta có thể dễ dàng nhận được thông tin của pokemon chúng ta muốn tìm kiếm. Ở đây ta có thể thấy số, chỉ số atk def cũng như các chỉ số tương tự khác của 1 con pokemon.

# Phần 4. Kết quả và thảo luận

**4.1. Các nguyên tắc thiết kế trực quan mà tôi đã áp dụng**

- Nguyên tắc Đơn giản hóa (Simplicity):

+ Lý do chọn lựa: Để tránh sự lộn xộn và làm cho giao diện người dùng dễ sử dụng. Một thiết kế đơn giản giúp người dùng dễ dàng tương tác với ứng dụng mà không bị rối mắt.

+ Áp dụng: Sử dụng ít màu sắc và hình ảnh, chỉ sử dụng các yếu tố cần thiết để truyền tải thông tin.

* Nguyên tắc Tương phản (Contrast):

+ Lý do chọn lựa: Tạo sự khác biệt rõ ràng giữa các yếu tố trên màn hình, giúp người dùng dễ dàng nhận biết và tập trung vào các phần quan trọng.

+ Áp dụng: Sử dụng màu sắc tương phản cho các phần tử chính và nền, sử dụng kiểu chữ đậm để nhấn mạnh các tiêu đề và nút bấm quan trọng.

* Nguyên tắc Nhất quán (Consistency):

+ Lý do chọn lựa: Đảm bảo rằng người dùng có thể dễ dàng làm quen và sử dụng ứng dụng một cách mượt mà.

+ Áp dụng: Giữ nguyên bố cục, kiểu chữ, và màu sắc cho tất cả các màn hình trong ứng dụng, sử dụng các biểu tượng và hình ảnh quen thuộc.

* Nguyên tắc Cân bằng (Balance):

+ Lý do chọn lựa: Đảm bảo rằng không có phần nào của giao diện nặng hơn hoặc nhẹ hơn phần khác, tạo cảm giác hài hòa và dễ nhìn.

+ Áp dụng: Phân bổ đều các yếu tố trên màn hình, sử dụng khoảng trắng hợp lý để tạo ra sự thoáng đãng.

* Nguyên tắc Định vị trực quan (Visual Hierarchy):

+ Lý do chọn lựa: Giúp người dùng dễ dàng nhận biết thông tin quan trọng và thứ tự ưu tiên của các yếu tố trên màn hình.

+ Áp dụng: Sử dụng kích thước, màu sắc và vị trí để tạo ra sự khác biệt giữa các yếu tố quan trọng và ít quan trọng hơn.

* Nguyên tắc Khả năng tiếp cận (Accessibility):

+ Lý do chọn lựa: Đảm bảo rằng tất cả người dùng, bao gồm cả những người có khuyết tật, có thể sử dụng ứng dụng một cách dễ dàng.

+ Áp dụng: Sử dụng kích thước chữ lớn, độ tương phản cao, và cung cấp các tùy chọn để thay đổi kích thước và màu sắc.

Bằng cách áp dụng những nguyên tắc này, tôi mong muốn tạo ra một giao diện người dùng không chỉ đẹp mắt mà còn dễ sử dụng và thân thiện với người dùng.

**4.2. Mô tả kết quả của bài tập lớn**

Về kết quả của bài tập lớn, tôi đã tạo ra các biểu đồ, đồ thị và trực quan hóa dữ liệu để phân tích và trình bày các thông tin quan trọng. Cụ thể:

* **Biểu đồ cột** được sử dụng để so sánh dữ liệu theo các nhóm hoặc categories khác nhau, giúp người xem dễ dàng nhận ra xu hướng và mối quan hệ giữa các yếu tố.
* **Biểu đồ hộp** được sử dụng để tổng quan các thông tin có trong tệp dữ liệu
* **Biểu đồ đường** được sử dụng để thể hiện diễn biến của dữ liệu , giúp người xem theo dõi được các biến động và xu hướng.
* **Biểu đồ tròn** được sử dụng để tổng quan các tộc hệ của nhân vật trong tệp dữ liệu, chúng ta có thể dễ dàng quan sát cái nào chiếm đa số, thiểu số.
* **Biểu đồ Sankey** được sử dụng để biết được lựa chọn tốt nhất cho đối thủ của mình
* **Biểu đồ PCA và KDE** được sử dụng để xem được sự phân bố của Pokemon khắp các thế giới, cho ta thấy được sự đa dạng của chúng .

**4.3. Thông tin nhận thấy từ trực quan hóa dữ liệu**

1. **Xu hướng và diễn biến**: Các biểu đồ đường và cột đã giúp tôi nhận ra các xu hướng và diễn biến của dữ liệu theo thời gian, từ đó có thể đưa ra các dự đoán và kế hoạch phù hợp.
2. **Sự phân bổ và tương quan**: Các biểu đồ tròn đã giúp tôi hiểu rõ hơn về sự phân bổ và tương quan giữa các thành phần của dữ liệu, từ đó có thể đưa ra các phân tích và kết luận hợp lý.
3. **So sánh và đối chiếu**: Các biểu đồ cột đã giúp tôi so sánh và đối chiếu dữ liệu giữa các nhóm hoặc categories, từ đó có thể nhận ra các điểm tương đồng và khác biệt, giúp đưa ra các kết luận chính xác hơn.

Tóm lại, việc áp dụng các nguyên tắc thiết kế trực quan và tạo ra các trực quan hóa dữ liệu phù hợp đã giúp tôi trình bày thông tin một cách rõ ràng, dễ hiểu và có ý nghĩa. Điều này đóng góp đáng kể vào việc phân tích và đưa ra các kết luận quan trọng từ dữ liệu.

# Phần 5. Tổng kết bài tập lớn

## 5.1. Kế hoạch

### *5.1.1. Kế hoạch dự kiến*

*Giai đoạn 1:*

Nhóm cùng nhau lựa chọn ý tưởng: tìm kiếm dữ liệu, phương pháp xử lý dữ liệu,… Cuối cùng nhóm thống nhất đề tài Pokemon, và áp dụng tài liệu mở để tăng thêm phần hiểu biết về thế giới

*Giai đoạn 2:*

**Phân chia công việc:**

|  |  |
| --- | --- |
| Lê Trung Kiên | Code trực quan hóa dữ liệu. Phân loại. Làm sạch dữ liệu. Phân tích dữ liệu. Phân tích mối tương quan. Tạo biểu đồ Viết báo cáo. làm slide. |
| Đinh Thái Phúc | Code trực quan hóa dữ liệu, Tìm dữ liệu, xử lý dữ liệu mở. Phân loại. Làm sạch dữ liệu. Phân tích dữ liệu Viết báo cáo. |

Bảng 3. Phân chia công việc

### *5.1.2. Phân chia công việc*

|  |  |  |  |
| --- | --- | --- | --- |
| Vai trò | | Lê Trung Kiên | Đinh Thái Phúc |
| Trưởng nhóm | Thành viên |
| STT | Nội dung công việc | Phần % công việc thực hiện | |
| 1 | Báo cáo BTL phần 1, 2 | 50% | 50% |
| 2 | Báo cáo BTL phần 3, 4, 5 | 50% | 50% |
| 3 | Trực quan hóa dữ liệu | 50% | 50% |
| 4 | Phân loại | 50% | 50% |
| 5 | Làm sạch dữ liệu | 50% | 50% |
| 6 | Phân tích dữ liệu | 50% | 50% |
| 7 | Phân tích mối tương quan | 50% | 50% |
| 8 | Làm slide | 100% | 0% |
| Tổng kết | | 100% | |

Bảng 4. Bảng nội dung công việc

## 5.2. Kết quả đạt được

Kết quả đạt được sau bài phân tích và trực quan hóa dữ liệu về Pokemon:

Hiểu rõ hơn về phân bố Pokemon trong thế giới : Phân tích đã cho thấy sự đa dạng lớn về tỷ lệ xuất hiện của pokemon thường, hiếm, huyền thoại, các tộc hệ rất đa dạng,…

Cho thấy sự hữu dụng của Pokedex giúp ta dễ dàng tìm thấy thông tin của Pokemon ta quan tâm đến

*Tóm lại*, bài phân tích này mang lại cái nhìn sâu sắc về sự phân bố và mối tương quan giữa các yếu tố liên quan đến Pokemon,.

## 5.3. Ưu điểm và hạn chế

### *5.3.1. Ưu điểm*

Trực quan hóa dữ liệu: Sử dụng biểu đồ và đồ thị giúp dễ dàng hiểu và phân tích dữ liệu một cách trực quan. Điều này giúp nhìn thấy mô hình và xu hướng một cách rõ ràng hơn, từ đó đưa ra những phân tích chính xác hơn.

Phân tích tương quan: Công cụ cho phép phân tích mối tương quan giữa các biến số, như tỷ lệ xuất hiện của Pokemon, từ đó giúp xác định các mối liên hệ và ảnh hưởng giữa chúng.

Khả năng tổng hợp thông tin: Phương pháp cho phép tổng hợp thông tin từ nhiều nguồn khác nhau vào một nền tảng trực quan và dễ tiếp cận, giúp cho việc đưa ra quyết định quản lý và phát triển có căn cứ khoa học.

Áp dụng linh hoạt: Công cụ và phương pháp có thể áp dụng linh hoạt cho nhiều lĩnh vực khác nhau, không chỉ giới hạn trong lĩnh vực Pokemon mà còn có thể mở rộng sang các lĩnh vực khác trong đời sống.

### *5.3.2. Hạn chế*

Giới hạn dữ liệu: Phương pháp phụ thuộc vào tính chính xác và sự đa dạng của dữ liệu sẵn có. Nếu dữ liệu không đủ lớn hoặc không đủ đại diện, các kết quả phân tích có thể không thể áp dụng rộng rãi.

Giới hạn mô hình hóa: Việc mô hình hóa dữ liệu dựa trên các biểu đồ và đồ thị có thể giới hạn trong việc đưa ra các dự đoán phức tạp hơn. Các mô hình phân tích này thường không đủ để dự đoán các xu hướng dài hạn hoặc phát triển tương lai.

Phụ thuộc vào người phân tích: Kết quả phân tích có thể phụ thuộc nhiều vào khả năng hiểu và phân tích của người sử dụng công cụ. Việc hiểu sai hoặc đưa ra những giả định sai lầm có thể dẫn đến kết quả không chính xác.

Cần sự hiểu biết chuyên môn: Việc sử dụng phương pháp và công cụ yêu cầu người phân tích có kiến thức vững về thống kê, phân tích dữ liệu và việc đọc hiểu các biểu đồ, đồ thị một cách chuyên sâu để có thể đưa ra các kết luận và nhận định phù hợp.

# Phần 6. Tài liệu tham khảo

[1]: Link github của giảng viên Nguyễn Văn Thiệu.

[thieu1995 (Nguyen Van Thieu) (github.com)](https://github.com/thieu1995)

[2]: Sample Data. Complete Pokémon Pokédex

[Pokémon Pokédex: list of Pokémon with stats | Pokémon Database (pokemondb.net)](https://pokemondb.net/pokedex/all)

[3]: Link Colab:

[Pokemon Index.ipynb - Colab (google.com)](https://colab.research.google.com/drive/1pI2no8wSvx07KM7xXjNM45p0tK__Zei2#scrollTo=MPNZFelUl0pl)

[4]: Link Github:

[Kien1804/trucquanhoadulieu (github.com)](https://github.com/Kien1804/trucquanhoadulieu)

[5]: Link Drive:

[DV-1-3-23(N07) - Google Drive](https://drive.google.com/drive/folders/1CYz5w6bohe31_KQCnx2jvGrKSELLjMHs)