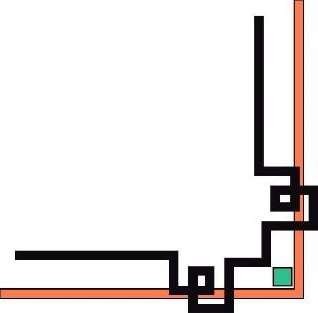
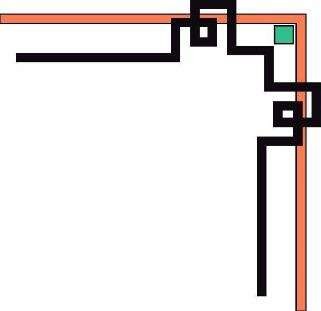
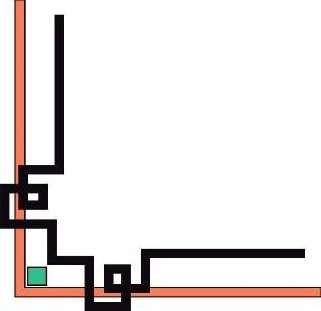
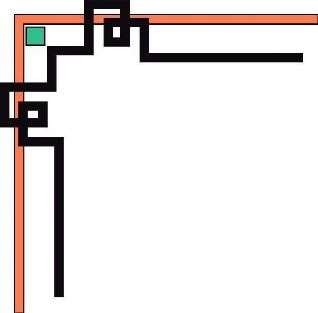
**ĐẠI HỌC HUẾ**



**KHOA K**Ỹ **THU**Ậ**T VÀ CÔNG NGH**Ệ

**BÁO CÁO ĐỒ ÁN**

**(hoặc TIỂU LUẬN, BÀI TẬP LỚN)**

**Học kỳ II, năm học 2022 - 2023**

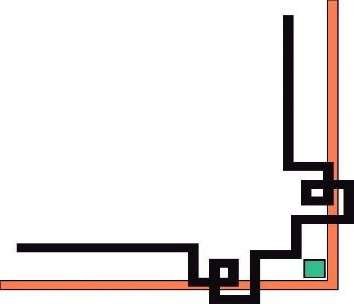
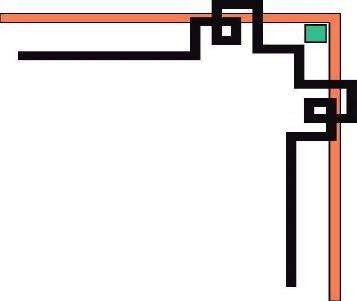
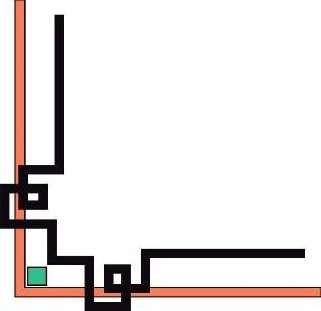
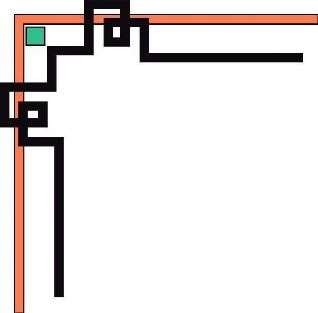
**Học phần:**

**Phân tích dữ liệu Python**

**Số phách**

*(Do hội đồng chấm thi ghi)*

**Thừa Thiên Huế, ngày 1 tháng 4 năm 2022**



**ĐẠI HỌC HUẾ**

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

**(Bìa phụ 2)**

**BÁO CÁO ĐỒ ÁN**

**(hoặc TIỂU LUẬN, BÀI TẬP LỚN)**

**Học kỳ II, năm học 2022 - 2023**

**Học phần:**

**Phân tích dữ liệu Python**

**Giảng viên hướng dẫn: Nguyễn Thế Dũng**

**Lớp: KHDL & TTNT K3**

**Sinh viên thực hiện: Trần Hậu Nguyên Đạt – Nhóm 8**

**Mã số sinh viên: 22E1020025**

*(ký tên và ghi rõ họ tên)*

**Số phách**

*(Do hội đồng chấm thi ghi)*

**Thừa Thiên Huế, ngày 1 tháng 4 năm 2023**

**ĐẠI HỌC HUẾ**

**KHOAN KỸ THUẬT VÀ CÔNG NGHỆ**

**PHIẾU ĐÁNH GIÁ ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN**

**Học kỳ II, năm học 2022-2023**

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1** | **Cán bộ chấm thi 2** |
| **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBTC 1:**  Bằng số:................................................  Bằng chữ:.............................................. | **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBTC :**  Bằng số:................................................  Bằng chữ:.............................................. |

Điểm kết luận: .....................................................................................................................

Bằng số: ...............................................................................................................................

Bằng chữ: ............................................................................................................................

Thừa Thiên Huế, ngày ....., tháng ....., năm 2023

**Cán bộ chấm thi 1 Cán bộ chấm thi 2**

*(Ký và ghi rõ họ tên) (Ký và ghi rõ họ tên)*

# LỜI CẢM ƠN

*Đầu tiên, em xin cảm ơn Đại học Huế đã cung cấp cho tôi môi trường học tập chuyên nghiệp và cơ hội để phát triển năng lực của mình.*

*Em cũng muốn gửi lời cảm ơn đến giáo viên hướng dẫn, Nguyễn Thế Dũng, người đã tận tình hỗ trợ và hướng dẫn trong suốt quá trình nghiên cứu và viết bài. Những kiến thức và kinh nghiệm mà ông chia sẻ đã giúp tôi hiểu sâu hơn về chủ đề tôi nghiên cứu.*

*Em cũng muốn cảm ơn gia đình và bạn bè của mình đã động viên và ủng hộ tôi trong suốt thời gian viết bài. Sự hỗ trợ của họ đã giúp tôi vượt qua những khó khăn và tìm được động lực để hoàn thành bài tiểu luận này.*

*Cuối cùng, em hy vọng rằng bài tiểu luận của em sẽ có ích và đóng góp vào lĩnh vực tôi nghiên cứu. Một lần nữa, em xin gửi lời cảm ơn tới tất cả những người đã hỗ trợ và động viên em trong quá trình này.*

*Em xin chân thành cảm ơn!*

# DANH MỤC HÌNH ẢNH

[Hình 1: Cú pháp đọc dữ liệu 9](#_Toc137567477)

[Hình 2: Kết quả khi đọc hết dữ liệu 10](#_Toc137567478)

[Hình 3: Kết quả khi lọc dữ liệu bằng loc 11](#_Toc137567479)

[Hình 4: Kết quả khi lọc dữ liệu bằng iloc 11](#_Toc137567480)

[Hình 5: Thống kê nhiệt độ cao nhất khu vực 12](#_Toc137567481)

[Hình 6: Thống kê nhiệt độ thấp nhất khu vực 12](#_Toc137567482)

[Hình 7: Thống kê nhiệt độ trung bình các năm dưới 2000 13](#_Toc137567483)

[Hình 8: Thông kê trung vị độ mặn từng khu vực 13](#_Toc137567484)

[Hình 9: Thống kê phương sai độ mặn của từng khu vực 14](#_Toc137567485)

[Hình 10: Thống kê phương sai hiệu chỉnh 15](#_Toc137567486)

[Hình 11: Thống kê độ lệch chuẩn độ mặn của từng khu vực 15](#_Toc137567487)

[Hình 12: Kết quả tứ phân vị 16](#_Toc137567488)

[Hình 13: : Bảng phân tổ nhiệt độ theo từng tháng trong cả nước 16](#_Toc137567489)

[Hình 14; Gộp nhóm cơ bản 17](#_Toc137567490)

[Hình 15: Dữ liệu biểu đồ tròn 19](#_Toc137567491)

[Hình 16; Biểu đồ tròn 19](#_Toc137567492)

[Hình 17: Dữ liệu biểu đồ cột 20](#_Toc137567493)

[Hình 18: Biểu đồ cột 20](#_Toc137567494)

[Hình 19: Dữ liệu biểu đồ đường 21](#_Toc137567495)

[Hình 20: Biểu đồ đường 21](#_Toc137567496)

[Hình 21: Dữ liệu biểu đồ Histogram 22](#_Toc137567497)

[Hình 22: Biểu đồ Histogram 22](#_Toc137567498)

[Hình 23: Dữ liệu cột ghép 23](#_Toc137567499)

[Hình 24: Biểu đồ cột ghép 23](#_Toc137567500)

[Hình 25: Dữ liệu biểu đồ cột chồng 24](#_Toc137567501)

[Hình 26: Biểu đồ cột chồng 24](#_Toc137567502)

[Hình 27: Dữ liệu biểu đồ phân tán 25](#_Toc137567503)

[Hình 28: Biểu đồ phân tán 25](#_Toc137567504)

[Hình 29: Tương quan giữa độ mặn và nhiệt độ 28](#_Toc137567505)

[Hình 30: Đồ thị biểu diễn tương quan giữa độ mặn và nhiệt độ 28](#_Toc137567506)

# DANH MỤC BẢNG BIỂU

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc137536742)

[DANH MỤC HÌNH ẢNH ii](#_Toc137536743)

[DANH MỤC BẢNG BIỂU iii](#_Toc137536744)

[MỤC LỤC iv](#_Toc137536745)

[CHƯƠNG 1: GIỚI THIỆU VỀ DỮ LIỆU 5](#_Toc137536746)

[1.1 Giới thiệu dữ liệu: 5](#_Toc137536747)

[1.2 Mô tả dữ liệu: 5](#_Toc137536748)

[CHƯƠNG 2: LÀM VIỆC VỚI DỮ LIỆU 6](#_Toc137536749)

[2.1 Đọc và lọc dữ liệu: 6](#_Toc137536750)

[2.1.1 Đọc dữ liệu 6](#_Toc137536751)

[2.1.2 Lọc dữ liệu 7](#_Toc137536752)

[2.1.3 Mã nguồn 7](#_Toc137536753)

[2.2 Mô tả dữ liệu 7](#_Toc137536754)

[2.2.1 Thống kê cơ bản dữ liệu 7](#_Toc137536755)

[2.2.2 Gộp nhóm cơ bản 11](#_Toc137536756)

[2.2.3 Một số biểu đồ về các bài toán trên 11](#_Toc137536757)

[2.2.4 Mã nguồn 14](#_Toc137536758)

[CHƯƠNG 3: PHÂN TÍCH CƠ BẢN CÁC BÀI TOÁN 1](#_Toc137536759)

[3.1 Bài toán ước lượng 1](#_Toc137536760)

[3.2 Bài toán kiểm định 1](#_Toc137536761)

[3.3 Bài toán tương quan 1](#_Toc137536762)

[3.4 Mã nguồn 3](#_Toc137536763)

[KẾT QUẢ KIẾM TRA ĐẠO VĂN 4](#_Toc137536764)

# CHƯƠNG 1: GIỚI THIỆU VỀ DỮ LIỆU

## 1.1 Giới thiệu dữ liệu:

Dữ liệu nhiệt độ, mùa, độ mặn và CHLFa (Chlorophyll a) trong một tháng và năm cung cấp thông tin về các yếu tố quan trọng liên quan đến môi trường và hệ sinh thái trong một khu vực cụ thể. Dưới đây là giới thiệu về mỗi yếu tố: Nhiệt độ: Dữ liệu nhiệt độ ghi lại mức độ nóng hoặc lạnh của môi trường trong tháng và năm nhất định. Nhiệt độ có thể được đo bằng độ Celsius (°C) hoặc độ Fahrenheit (°F). Dữ liệu nhiệt độ cung cấp thông tin về xu hướng nhiệt độ hàng ngày, biên độ nhiệt độ, và trung bình nhiệt độ trong một khoảng thời gian xác định. Mùa: Dữ liệu mùa xác định mùa trong một tháng và năm cụ thể. Mùa bao gồm các khái niệm như mùa xuân, mùa hạ, mùa thu và mùa đông. Thông qua dữ liệu mùa, ta có thể xác định được thời điểm mà một tháng cụ thể thuộc về mùa nào, điều này có thể ảnh hưởng đến các quá trình tự nhiên và hoạt động của con người trong khu vực đó. Độ mặn: Dữ liệu độ mặn đo lường lượng muối có trong một môi trường nước cụ thể. Độ mặn thường được đo bằng các đơn vị như phần/milimét (ppt/pptm), phần/trăm (%), hay gram muối trong một lít nước (g/L). Độ mặn có thể ảnh hưởng đến sự sống của các sinh vật thủy sinh và cũng có liên quan đến quy trình thủy văn trong hệ sinh thái. CHLFa (Chlorophyll a): Dữ liệu CHLFa cho biết nồng độ Chlorophyll a trong một môi trường nước. Chlorophyll a là một chất quang hợp chính trong các tế bào thực vật và các sinh vật thủy sinh khác. Nồng độ CHLFa có thể cung cấp thông tin về mức độ sản xuất sinh học và tình trạng sức khỏe của hệ thống sinh thái nước. Các dữ liệu này thường được thu thập từ các trạm quan trắc môi trường, các đài quan trắc hoặc các nghiên cứu khoa học. Việc theo dõi và phân tích dữ liệu này giúp hiểu rõ hơn về sự biến đổi của môi trường nước và hệ sinh thái, từ đó đưa ra các quyết định và biện pháp bảo vệ và quản lý môi trường hiệu quả

## 1.2 Mô tả dữ liệu:

A screenshot of a computer screen

Description automatically generated with medium confidence

Trong dữ liệu về mùa, độ mặn, nhiệt độ, CHLFa tổng cộng có 6 cột bao gồm cột năm, cột tháng, cột mùa, cột độ mặn, cột nhiệt độ, cột CHFLa, Cột khu vực. Trong dữ liệu còn có tổng cộng 4420 dòng là thông tin về mùa, độ mặn, nhiệt độ, CHLFa trong tháng bất kỳ từ năm 1990 đến năm 2005. Từ các thông tin mà dữ liệu mang lại nhóm em đã đưa vào làm dữ liệu của nhóm dùng để giải các bài tập cho môn học phân tích dữ liệu Python.

* Year: Dữ liệu được khảo sát từ năm 1990 cho đến năm 2005
* Month: Dữ liệu được thu thập qua đầy đủ 12 tháng của các năm.
* Season: Dữ liệu thu nhập gồm đủ 4 mùa xuân, hạ, thu, đông.
* Salinity: độ mặn của nước được đo bằng ppt (phần trăm nghìn) có giá trị nằm trong khoảng từ 4,69 đến 34,5 g/kg.
* Temperature: nhiệt độ được đo bằng độ C (°C) với dữ liệu khảo sát trên, giá trị của nhiệt độ dao động từ -0,83 đến 23,84°C.
* CHLFa: nồng độ của một hợp chất hóa học cụ thể, được đo bằng mg/L, nồng độ trong dữ liệu được đo từ 0,42 đến 82,9 mg/L.

# CHƯƠNG 2: LÀM VIỆC VỚI DỮ LIỆU

## 2.1 Đọc và lọc dữ liệu:

### 2.1.1 Đọc dữ liệu

Để đọc và lọc dữ liệu trong Python, nhóm em sử dụng các thư viện như Pandas và NumPy.

Cú pháp python dùng để đọc dữ liệu bằng thư viện pandas:

**df = pd.read\_excel('path\_to\_excel\_file.xlsx')**

Trong đó:

**Pd:** là viết tắt của pandas, đã được import từ thư viện Pandas.

**read\_excel:** là phương thức của Pandas được sử dụng để đọc file Excel.

**'path\_to\_excel\_file.xlsx':** là đường dẫn đến file Excel mà bạn muốn đọc. Bạn có thể cung cấp đường dẫn tuyệt đối hoặc đường dẫn tương đối của file Excel.

Ngoài ra, phương thức **read\_excel** cung cấp nhiều tham số tùy chọn để tùy chỉnh quá trình đọc file Excel, chẳng hạn như **sheet\_name** để chọn sheet cụ thể, **header** để chỉ định hàng đầu tiên làm header, và nhiều tham số khác.

Ví dụ:

**A picture containing text, font, screenshot, white

Description automatically generated**

Hình 1: Cú pháp đọc dữ liệu

Và khi in kết quả ra màn hình:

A screenshot of a computer screen

Description automatically generated with medium confidence

Hình 2: Kết quả khi đọc hết dữ liệu

### 2.1.2 Lọc dữ liệu

Để lọc dữ liệu nhóm em sử dụng loc và iloc để lọc hàng hoặc cột

Cú pháp chung:

**df.loc[row\_label, column\_label]**

Trong đó:

* **Df**: là DataFrame mà bạn muốn truy cập.
* **row\_label:** là nhãn (label) hoặc danh sách nhãn của các hàng (rows) bạn muốn lấy.
* **column\_label:** là nhãn (label) hoặc danh sách nhãn của các cột (columns) bạn muốn lấy.

**df.iloc[row\_index, column\_index]**

Trong đó:

* **df:** là DataFrame hoặc Series mà bạn muốn truy cập.
* **row\_index:** là chỉ số (index) hoặc danh sách chỉ số của các hàng (rows) bạn muốn lấy.
* **column\_index:** là chỉ số (index) hoặc danh sách chỉ số của các cột (columns) bạn muốn lấy.

Ví dụ khi dùng loc: Lọc dữ liệu từ hàng 1 tới hàng 5 của cột Year và cột Season

Cú pháp:

print(df.loc[1:5, [**"Year"**, **"Season"**]])

Sau khi thực hiện lệnh ta có kết quả như sau:

A picture containing text, font, screenshot

Description automatically generated

Hình 3: Kết quả khi lọc dữ liệu bằng loc

Ví dụ khi dùng iloc: Lọc dữ liệu từ hàng 1 tới hàng 5 và cột 0 tới cột 4 (cột hoặc hàng đầu tiên sẽ là số 0).

Cú pháp: print(df.iloc[1:5, 0:4])

A picture containing text, font, screenshot, number

Description automatically generated

Hình 4: Kết quả khi lọc dữ liệu bằng iloc

### 2.1.3 Mã nguồn

Mã nguồn python về các bài toán trên:

<https://github.com/phanbahung1124/motadulieu/blob/master/motadulieu.py>

## 2.2 Mô tả dữ liệu

### 2.2.1 Thống kê cơ bản dữ liệu

**Max**: Giá trị lớn nhất

Ví dụ: Thống kê nhiệt độ cao nhất của từng khu vực

A screenshot of a computer screen

Description automatically generated with low confidence

Hình 5: Thống kê nhiệt độ cao nhất khu vực

**Min**: Giá trị nhỏ nhất

Ví dụ: Thống kê nhiệt độ thấp nhất của từng khu vực

A screenshot of a computer

Description automatically generated with low confidence

Hình 6: Thống kê nhiệt độ thấp nhất khu vực

**Mean**: Giá trị trung bình

Ví dụ: Thống kê nhiệt độ trung bình các năm dưới năm 2000

A picture containing text, screenshot, font, number

Description automatically generated

Hình 7: Thống kê nhiệt độ trung bình các năm dưới 2000

**Median**: Trung vị một khái niệm thống kê dùng để đại diện cho giá trị trung tâm của một tập dữ liệu số. Nó được xác định bằng cách sắp xếp các giá trị trong tập dữ liệu theo thứ tự tăng dần hoặc giảm dần và chọn giá trị ở vị trí giữa.

Ví dụ: Thống kê trung vị độ mặn của từng khu vực

A picture containing text, screenshot, font, number

Description automatically generated

Hình 8: Thông kê trung vị độ mặn từng khu vực

**Phương sai**: là một khái niệm trong thống kê dùng để đo độ biến thiên của một tập dữ liệu. Nó cho biết mức độ phân tán của các giá trị trong tập dữ liệu so với giá trị trung bình. Phương sai được tính bằng cách tính trung bình của bình phương hiệu giữa mỗi giá trị trong tập dữ liệu và giá trị trung bình của tập dữ liệu. Công thức tính phương sai của một tập dữ liệu gồm n phần tử là:

Trong đó:

* **xi** là giá trị của phần tử thứ i trong tập dữ liệu.
* **x̄** là giá trị trung bình của tập dữ liệu.
* **Σ** là ký hiệu tổng của tất cả các phần tử trong tập dữ liệu.

Vi dụ: Thống kê phương sai độ mặn của từng khu vực

A screenshot of a computer

Description automatically generated with low confidence

Hình 9: Thống kê phương sai độ mặn của từng khu vực

**Phương sai hiệu chỉnh**: là một biến thể của phương sai được sử dụng để ước lượng phương sai trong mẫu dữ liệu. Nó được sử dụng khi ta muốn điều chỉnh phương sai để đánh giá mức độ biến thiên của quần thể ban đầu, dựa trên một mẫu nhỏ hơn của dữ liệu.

Phương sai hiệu chỉnh được tính bằng cách sử dụng một đại lượng điều chỉnh gọi là hệ số hiệu chỉnh (correction factor). Hệ số hiệu chỉnh là một chỉ số phụ thuộc vào kích thước mẫu, và nó được sử dụng để điều chỉnh phương sai ước lượng để tương thích với quần thể ban đầu.

Công thức tính phương sai hiệu chỉnh là:

Trong đó:

* **n** là kích thước mẫu dữ liệu.
* **Phương sai** là phương sai ước lượng được tính bằng công thức thông thường.

Ví dụ: Thống kê phương sai hiệu chỉnh của độ mặn, nhiệt độ, CHLFa có kết quả như sau:

A picture containing text, font, screenshot, white

Description automatically generated

Hình 10: Thống kê phương sai hiệu chỉnh

Độ lệch chuẩn: một khái niệm trong thống kê dùng để đo mức độ phân tán của các giá trị trong một tập dữ liệu. Nó là căn bậc hai của phương sai và đơn vị đo của nó giống với đơn vị đo của tập dữ liệu ban đầu.

Trong đó:

* **xi** là giá trị của phần tử thứ i trong tập dữ liệu.
* **x̄** là giá trị trung bình của tập dữ liệu.
* **Σ** là ký hiệu tổng của tất cả các phần tử trong tập dữ liệu.

Ví dụ: Thống kê độ lệch chuẩn độ mặn của từng khu vực

A screenshot of a computer

Description automatically generated with low confidence

Hình 11: Thống kê độ lệch chuẩn độ mặn của từng khu vực

**Tứ phân vị**: một bảng dùng để tổ chức và hiển thị thông tin về tần suất xuất hiện của các giá trị trong một tập dữ liệu. Nó thường được sử dụng để tổng hợp dữ liệu rời rạc thành các nhóm và đếm số lượng phần tử trong mỗi nhóm.

Ví dụ: ta cho 1 tập dữ liệu data = [12, 15, 17, 20, 22, 24, 25, 27, 30, 32] thì kết quả tứ phân vị của tập dữ liệu sẽ là:

A picture containing font, screenshot, white, text

Description automatically generated

Hình 12: Kết quả tứ phân vị

**Bảng phân tổ**: là một bảng dùng để tổ chức và hiển thị thông tin về tần suất xuất hiện của các giá trị trong một tập dữ liệu. Nó thường được sử dụng để tổng hợp dữ liệu rời rạc thành các nhóm và đếm số lượng phần tử trong mỗi nhóm.

Ví dụ: Bảng phân tổ nhiệt độ theo từng tháng trong cả nước với các khoảng nhiệt độ như sau:

A picture containing text, screenshot, number, font

Description automatically generated

Hình 13: : Bảng phân tổ nhiệt độ theo từng tháng trong cả nước

### 2.2.2 Gộp nhóm cơ bản

Về gộp nhóm cơ bản là ta sẽ gộp 2 hoặc nhiều cột dữ liệu để thống kê dữ liệu đơn giản hơn. Để sử dụng gộp nhóm trong python ta sử dụng lệnh groubpy trong thư viện pandas.

Cú pháp chung:

**df.groupby(by, axis=0, level=None, as\_index=True, sort=True, group\_keys=True, squeeze=False, observed=False)**

Trong đó:

* **by**: Một đối tượng hoặc tên cột để gộp nhóm theo. Nó có thể là một chuỗi, danh sách, hoặc một hàm ánh xạ.
* **axis**: Trục dọc (0) hoặc trục ngang (1) để gộp nhóm. Mặc định là 0.
* **level**: Nếu DataFrame có các cấu trúc chỉ mục đa cấp, level xác định cấp độ chỉ mục để gộp nhóm.
* **as\_index**: Xác định liệu có sử dụng cột gộp nhóm làm chỉ mục của kết quả gộp hay không. Mặc định là True.
* **sort**: Xác định liệu kết quả gộp có được sắp xếp dựa trên các cột gộp hay không. Mặc định là True.
* **group\_keys**: Xác định liệu chỉ mục gộp có được thêm vào kết quả gộp hay không. Mặc định là True.
* **squeeze**: Xác định liệu kết quả gộp có được nén thành Series nếu có thể hay không. Mặc định là False.
* **observed**: Xác định liệu chỉ mục gộp có được sắp xếp dựa trên các giá trị quan sát đã có hay không. Mặc định là False.

Sau khi áp dụng groupby, bạn có thể sử dụng các phương thức như mean(), sum(), count(), min(), max() v.v. để tính toán hoặc thống kê trên từng nhóm.

Ví dụ: Tìm nhiệt độ cao nhất của từng khu vực, đây ta chỉ quan tâm 2 cột đó là nhiệt độ và khu vực.

Mã lệnh python:

dl = df[[**"Season"**, **"Temperature"**]]  
dl.max = df.groupby(**"Season"**).mean()

print(dl.max)

Và đây là kết quả sau khi thực hiện lệnh:

A screenshot of a computer screen

Description automatically generated with low confidence

Hình 14; Gộp nhóm cơ bản

### 2.2.3 Một số biểu đồ về các bài toán trên

Biểu đồ là một công cụ mạnh mẽ trong việc trực quan hóa dữ liệu. Chúng giúp chúng ta hiểu và trình bày thông tin một cách dễ hiểu và hấp dẫn. Biểu đồ có thể được sử dụng trong nhiều lĩnh vực khác nhau như khoa học, kinh doanh, thống kê, truyền thông và giáo dục.

Dưới đây là một số loại biểu đồ phổ biến:

* Biểu đồ cột (bar chart): Biểu đồ này sử dụng các cột để biểu diễn giá trị của các biến và so sánh chúng. Thích hợp để so sánh các giá trị hoặc phân loại.
* Biểu đồ đường (line chart): Biểu đồ này sử dụng các đường nối để biểu diễn sự thay đổi của một biến theo thời gian hoặc các giá trị liên tục khác.
* Biểu đồ hình tròn (pie chart): Biểu đồ này sử dụng các phần tương ứng với tỉ lệ phần trăm để biểu diễn phân phối của một biến theo các nhóm.

Ngoài ra ta có thể sử dụng thêm biểu đồ mới bằng cách kết hợp các biểu đồ trên như cột ghép, cột chồng, cột kết hợp đường, …

Thư viện phổ biến như Matplotlib, Seaborn và Plotly trong Python cung cấp nhiều chức năng và công cụ để tạo ra các biểu đồ đa dạng và tùy chỉnh. Để vẽ một biểu đồ trong Python, bạn có thể thực hiện các bước sau:

* Import các thư viện: Bắt đầu bằng việc import các thư viện cần thiết để vẽ biểu đồ. Một số thư viện phổ biến là Matplotlib, Seaborn, Plotly và Pandas.
* Chuẩn bị dữ liệu: Đảm bảo rằng bạn đã có dữ liệu cần thiết để vẽ biểu đồ. Dữ liệu có thể được lưu trữ trong các cấu trúc dữ liệu như danh sách, mảng, DataFrame hoặc từ các nguồn dữ liệu khác như tệp CSV, cơ sở dữ liệu, API, vv.
* Tạo đối tượng biểu đồ: Sử dụng các hàm và phương thức của thư viện để tạo đối tượng biểu đồ. Ví dụ: plt.figure() trong Matplotlib, sns.Figure() trong Seaborn hoặc px.Figure() trong Plotly.
* Tạo biểu đồ: Gọi các hàm và phương thức để tạo biểu đồ với dữ liệu đã chuẩn bị. Ví dụ: plt.plot() trong Matplotlib, sns.barplot() trong Seaborn, px.line() trong Plotly, vv.
* Tùy chỉnh biểu đồ: Sử dụng các phương thức và tham số của đối tượng biểu đồ để tùy chỉnh các yếu tố của biểu đồ như tiêu đề, nhãn trục, màu sắc, kích thước và phong cách.
* Hiển thị biểu đồ: Gọi phương thức hiển thị để hiển thị biểu đồ đã tạo ra.

**Biểu đồ tròn:**

Biểu đồ tròn (pie chart) là một loại biểu đồ trực quan được sử dụng để biểu diễn phân phối tỉ lệ phần trăm của các mục trong một tập dữ liệu. Nó thể hiện cách mà một toàn bộ hoặc một phần được chia thành các phần tương ứng với tỷ lệ phần trăm của chúng.

Ví dụ: Biểu đồ thể hiện nhiệt độ khu WZ năm 1990-2005

Để giải quyết yêu cầu bài toán đó, cần chuẩn bị bảng dữ liệu phù hợp: tính toán tổng nhiệt độ từ năm 1990-2005 trong tập dữ liệu. Phương thức được sử dụng là groupby(), tiến hành gộp nhóm tháng và khu vực để tính toán tổng trên cột dữ liệu “Temperature”.

Bảng dữ liệu được chuẩn bị bởi Python trả về kết quả như sau:

A picture containing text, font, white, receipt

Description automatically generated

Hình 15: Dữ liệu biểu đồ tròn

Dựa vào bảng dữ liệu trên ta vẽ được biểu đồ hình tròn như sau:

A picture containing text, screenshot, diagram, font

Description automatically generated

Hình 16; Biểu đồ tròn

**Biểu đồ cột:**

Biểu đồ cột (bar chart) là một loại biểu đồ trực quan được sử dụng để biểu diễn và so sánh giá trị của các mục trong một tập dữ liệu. Nó thể hiện mức độ khác biệt giữa các giá trị và giúp người đọc nhận ra các xu hướng, so sánh tương quan và phân loại dữ liệu.

Ví dụ: Thống kê nhiệt độ khu vực GM trong năm 2005

Để giải quyết yêu cầu bài toán đó, cần chuẩn bị bảng dữ liệu phù hợp: tính toán tổng nhiệt độ năm 2005 trong tập dữ liệu. Phương thức được sử dụng là groupby(), tiến hành gộp nhóm tháng và dữ liệu trong khu vực GM của năm 2005 để tính toán tổng trên cột dữ liệu “Temperature”.

Bảng dữ liệu được chuẩn bị bởi Python trả về kết quả như sau:

A picture containing text, screenshot, font

Description automatically generated

Hình 17: Dữ liệu biểu đồ cột

Dựa vào bảng dữ liệu trên ta vẽ được biểu đồ cột như sau:

A picture containing text, screenshot, plot, font

Description automatically generated

Hình 18: Biểu đồ cột

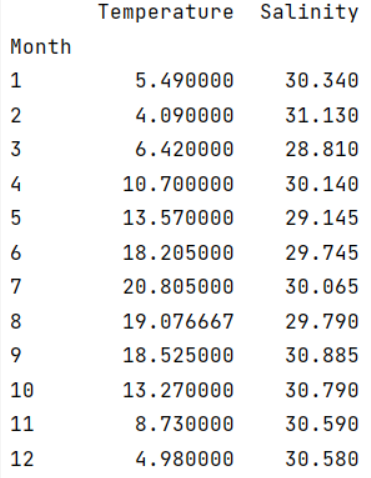
**Biểu đồ đường:**

Biểu đồ đường (line chart) là một loại biểu đồ trực quan được sử dụng để biểu diễn sự thay đổi của một biến hoặc một tập hợp các biến theo thời gian hoặc các điểm dữ liệu liên tục. Nó thể hiện xu hướng, mức độ biến đổi và mối tương quan giữa các giá trị.

Ví dụ: Thống kê sự thay đổi độ mặn khu GM trong những tháng năm 2005

Để giải quyết yêu cầu bài toán đó, cần chuẩn bị bảng dữ liệu phù hợp: tính toán độ mặn qua những tháng trong năm 2005 trong tập dữ liệu . Phương thức được sử dụng là groupby(), tiến hành gộp nhóm tháng và dữ liệu trong khu vực GM của năm 2005 để tính toán trên cột dữ liệu “Salinity”.

Bảng dữ liệu được chuẩn bị bởi Python trả về kết quả như sau:



Hình 19: Dữ liệu biểu đồ đường

Dựa vào bảng dữ liệu trên ta vẽ được biểu đồ đường như sau:

A picture containing text, diagram, plot, line

Description automatically generated

Hình 20: Biểu đồ đường

**Biểu đồ Histogram:**

Biểu đồ Histogram là một loại biểu đồ thống kê được sử dụng để biểu diễn phân phối của một biến số. Nó chia dữ liệu thành các khoảng (các bin) và đếm số lượng quan sát rơi vào mỗi khoảng. Biểu đồ Histogram thể hiện tần suất xuất hiện của các giá trị trong dữ liệu và giúp hiểu rõ hơn về phân phối của biến.\

Để giải quyết yêu cầu bài toán đó, cần chuẩn bị bảng dữ liệu phù hợp: tính toán nhiệt độ qua những tháng trong tập dữ liệu. Phương thức được sử dụng là groupby(), tiến hành gộp nhóm tháng và dữ liệu trong khu vực ED tính toán trên cột dữ liệu “Temperature”.

Bảng dữ liệu được chuẩn bị bởi Python trả về kết quả như sau:

A picture containing text, screenshot, font, number

Description automatically generated

Hình 21: Dữ liệu biểu đồ Histogram

Dựa vào bảng dữ liệu trên ta vẽ được biểu đồ Histogram như sau:

A picture containing text, diagram, screenshot, font

Description automatically generated

Hình 22: Biểu đồ Histogram

**Biểu đồ cột ghép:**

Biểu đồ cột ghép (grouped bar chart) là một loại biểu đồ sử dụng để so sánh giá trị của các mục trong các nhóm khác nhau. Nó cho phép so sánh một biến theo nhiều nhóm và đồng thời so sánh giá trị của các mục trong cùng một nhóm.

Để giải quyết yêu cầu bài toán đó, cần chuẩn bị bảng dữ liệu phù hợp: tính toán nhiệt độ và độ mặn qua những tháng của năm 2001 khu vực VD trong tập dữ liệu. Phương thức được sử dụng là groupby(), tiến hành gộp nhóm tháng và dữ liệu trong năm 2001 khu vực VD tính toán trên cột dữ liệu “Temperature” và “Salinity”.

Bảng dữ liệu được chuẩn bị bởi Python trả về kết quả như sau:

A picture containing text, screenshot, font, number

Description automatically generated

Hình 23: Dữ liệu cột ghép

Dựa vào bảng dữ liệu trên ta vẽ được biểu đồ cột ghép như sau:

A picture containing text, screenshot, font, parallel

Description automatically generated

Hình 24: Biểu đồ cột ghép

**Biểu đồ cột chồng:**

Biểu đồ cột chồng (stacked bar chart) là một loại biểu đồ được sử dụng để biểu diễn sự phân phối của một biến trong các nhóm khác nhau và tổng của các giá trị trong mỗi nhóm. Nó cho phép so sánh tổng của các giá trị trong từng nhóm và đồng thời biểu thị phần trăm đóng góp của các mục riêng lẻ trong tổng.

Để giải quyết yêu cầu bài toán đó, cần chuẩn bị bảng dữ liệu phù hợp: tính toán nhiệt độ và độ mặn qua những tháng của năm 2001 khu vực VD trong tập dữ liệu. Phương thức được sử dụng là groupby(), tiến hành gộp nhóm tháng và dữ liệu trong năm 2001 khu vực VD tính toán trên cột dữ liệu “Temperature” và “Salinity”.

Bảng dữ liệu được chuẩn bị bởi Python trả về kết quả như sau:

A picture containing text, screenshot, font, number

Description automatically generated

Hình 25: Dữ liệu biểu đồ cột chồng

Dựa vào bảng dữ liệu trên ta vẽ được biểu đồ cột chồng như sau:

A picture containing text, screenshot, font, plot

Description automatically generated

Hình 26: Biểu đồ cột chồng

**Biểu đồ phân tán:**

Biểu đồ phân tán (scatter plot) là một loại biểu đồ sử dụng để biểu diễn mối quan hệ giữa hai biến. Nó cho phép xem xét sự phân bố và mối liên hệ giữa các giá trị của hai biến số.

Để giải quyết yêu cầu bài toán đó, cần chuẩn bị bảng dữ liệu phù hợp: tính mối tương quan giữa nhiệt độ và độ mặn cảu khu vực VD trong tập dữ liệu.

Bảng dữ liệu được chuẩn bị bởi Python trả về kết quả như sau:

A screenshot of a computer screen

Description automatically generated with low confidence

Hình 27: Dữ liệu biểu đồ phân tán

Dựa vào bảng dữ liệu trên ta vẽ được biểu đồ phân tán như sau:

A picture containing screenshot

Description automatically generated

Hình 28: Biểu đồ phân tán

### 2.2.4 Mã nguồn

Mã nguồn python giải các bài toán trên như sau:

Thông kê và gộp nhóm cơ bản với dữ liệu:

<https://github.com/TranDucDuyAnh/python_data_analyse/blob/master/basicstat.py>

Đồ thị:

<https://github.com/TranDucDuyAnh/python_data_analyse/blob/master/plot.py>

# CHƯƠNG 3: PHÂN TÍCH CƠ BẢN CÁC BÀI TOÁN

## 3.1 Bài toán ước lượng

Để khảo sát độ mặn theo dữ liệu của nhóm qua nhiều tháng quan sát ta có kết quả như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X(Độ mặn) | 4-6 | 6-8 | 8-10 | 10-12 | 12-14 |
| Tháng | 11 | 7 | 22 | 35 | 20 |

1. Từ bản phân phối này ước lượng tỉ lệ độ mặn lớn hơn 24 với độ tin cậy 90%
2. Với độ lệch chuẩn xây dựng khoảng tin cậy 90% cho độ mặn trung bình với dữ liệu trên.

Với câu a) khi giải bằng python ta có kết quả như sau:



Với câu b) khi giải bằng python ta có kết quả như sau:



## 3.2 Bài toán kiểm định

**Kiểm định kì vọng bằng T-TEST**

Bài toán: Nhiệt độ trung bình thông thường ở khu OS là 12 độ. Kiểm định giả thuyết với mức ý nghĩa là 5%.

**Kiểm định tỉ lệ bằng Z-TEST**

Bài toán: Từ 1990 đến 2005 có 50% tháng quan sát mà nồng độ CHLFa ở khu WS lớn hơn 5. Kiểm định giả thuyết với mức ý nghĩa là 5%.

**Kiểm định phương sai bằng ANOVA TEST**

Bài toán: Từ năm 2000 trở về sau, độ mặn trung bình ở 3 khu vực VD, ED, GM là như nhau với mức ý nghĩa 5%.

Kết quả của các bài toán khi giải bằng python như sau:

A picture containing text, screenshot, font, algebra

Description automatically generated

## 3.3 Bài toán tương quan

Bài toán: khu vực VD để xét tương quan giữa độ mặn và nhiệt độ

Kết quả bài toán trên khi giải bằng python như sau:

A screenshot of a computer screen

Description automatically generated with low confidence

Hình 29: Tương quan giữa độ mặn và nhiệt độ

Ta có thể quan sát thêm qua biểu đồ sau:

A picture containing screenshot

Description automatically generated

Hình 30: Đồ thị biểu diễn tương quan giữa độ mặn và nhiệt độ

## 3.4 Mã nguồn

Mã nguồn python các bài toán trên:

Ước lượng:

<https://github.com/trandat2509/uoc-luong/blob/main/uoc_luong.py>

Kiểm định:

<https://github.com/TranDucDuyAnh/python_data_analyse/blob/master/test.py>

Tương quan:

<https://github.com/TranDucDuyAnh/python_data_analyse/blob/master/correlation.py>

# TÀI LIỆU THAM KHẢO

# KẾT QUẢ KIỂM TRA ĐẠO VĂN

A screenshot of a phone

Description automatically generated with low confidence