

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ  
BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**THỰC TẬP ĐỒ ÁN CƠ SỞ NGÀNH  
HỌC KỲ 1, NĂM HỌC 2023 – 2024  
TÌM HIỂU CÔNG CỤ AWS VÀ ỨNG DỤNG TRỰC QUAN HÓA DỮ LIỆU**

Giáo viên hướng dẫn:  
ThS. Nguyễn Khắc Quốc

Sinh viên thực hiện:  
Họ tên: Nguyễn Hữu Trí  
MSSV: 110121120  
Lớp: DA21TTB

*Trà Vinh, tháng..... năm.....*

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ  
BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**THỰC TẬP ĐỒ ÁN CƠ SỞ NGÀNH  
HỌC KỲ 1, NĂM HỌC 2023 – 2024  
TÌM HIỂU CÔNG CỤ AWS VÀ ỨNG DỤNG TRỰC QUAN HÓA DỮ LIỆU**

Giáo viên hướng dẫn:  
ThS. Nguyễn Khắc Quốc

Sinh viên thực hiện:  
Họ tên: Nguyễn Hữu Trí  
MSSV: 110121120  
Lớp: DA21TTB

*Trà Vinh, tháng ..... năm .....*

### NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

[illegible]

Trà Vinh, ngày ..... tháng ..... năm .....  
**Giáo viên hướng dẫn**  
*(Ký tên và ghi rõ họ tên)*

## NHẬN XÉT CỦA THÀNH VIÊN HỘI ĐỒNG

This image shows a full page of a document template designed for handwriting practice or general note-taking. It consists of approximately 28 evenly spaced horizontal dotted lines across the entire width of the page. The background is plain white, and there are no margins, headers, footers, or other markings present.

Trà Vinh, ngày ..... tháng ..... năm .....

**Thành viên hội đồng**  
(Ký tên và ghi rõ họ tên)

## **LỜI CẢM ƠN**

Em xin chân thành cảm ơn thầy ThS.Nguyễn Khắc Quốc đã dành thời gian hướng dẫn tạo điều kiện và giúp em có thêm kiến thức trong quá trình thực hiện đề tài.

Thông qua quá trình thực hiện đề tài, em đã phần nào tích lũy được những kiến thức về trực quan hóa dữ liệu.

Mặc dù đã cố gắng hết sức trong quá trình học tập và thực hiện đề tài nhưng em cũng không thể tránh khỏi những sai sót, em rất mong nhận được sự góp ý và thông cảm của quý thầy cô, giúp em rút kinh nghiệm cho những dự án trong tương lai.

Em xin chân thành cảm ơn.

**MỤC LỤC**

<b>MỞ ĐẦU.....</b>	<b>9</b>
<b>CHƯƠNG 1: TỔNG QUAN.....</b>	<b>11</b>
<b>CHƯƠNG 2: NGHIÊN CỨU LÝ THUYẾT.....</b>	<b>12</b>
<b>2.1 Trục quan hóa dữ liệu .....</b>	<b>12</b>
2.1.1 Khái niệm .....	12
2.1.2 Các thành phần trục quan hóa dữ liệu.....	12
2.1.3 Các loại biểu đồ.....	13
2.1.4 Các bước trục quan hóa dữ liệu .....	16
2.1.5 Ứng dụng.....	17
<b>2.2 Amazon Web Service.....</b>	<b>18</b>
2.2.1 Khái niệm .....	18
2.2.2 Amazon QuickSight.....	18
2.2.3 Amazon SageMaker .....	18
2.2.4 Amazon Athena.....	19
2.2.5. Amazon Kinesis .....	19
2.2.6 Amazon EMR.....	20
<b>CHƯƠNG 3: HIỆN THỰC HÓA NGHIÊN CỨU .....</b>	<b>21</b>
<b>3.1 Áp dụng trục quan hóa dữ liệu trên AWS .....</b>	<b>21</b>
3.1.1 Giao diện chính của công cụ .....	21
3.1.2 Tạo biểu đồ trục quan hóa.....	24
<b>CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU.....</b>	<b>28</b>
<b>CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....</b>	<b>29</b>
<b>DANH MỤC TÀI LIỆU THAM KHẢO .....</b>	<b>30</b>
<b>PHỤ LỤC .....</b>	<b>31</b>

**DANH MỤC HÌNH ẢNH**

Hình 2. 1 Line chart.....	13
Hình 2. 2 Bar chart .....	14
Hình 2. 3 Pie chart .....	14
Hình 2. 4 Scatter plot.....	15
Hình 2. 5 Box plot .....	15
Hình 2. 6 Map.....	16
Hình 3. 1 Giao diện chính của công cụ AWS .....	21
Hình 3. 2 Giao diện tìm và truy cập dịch vụ AWS QuickSight .....	21
Hình 3. 3 Giao diện tạo dataset AWS QuickSight .....	22
Hình 3. 4 Giao diện tìm và truy cập dịch vụ AWS QuickSight .....	22
Hình 3. 5 Giao diện tập dữ liệu được chọn AWS QuickSight .....	23
Hình 3. 6 Giao diện tìm và truy cập dịch vụ AWS QuickSight .....	24
Hình 3. 7 Giao diện tìm và truy cập dịch vụ AWS QuickSight .....	24
Hình 3. 8 Chức năng Visuals.....	25
Hình 3. 9 Biểu đồ Phân Phối Giá của Các Nhà Sản Xuất .....	26
Hình 3. 10 Mối Quan Hệ giữa Kích Thước Màn Hình và Giá theo Tên loại.....	26
Hình 3. 11 Biểu đồ so sánh trọng lượng trung bình của các loại laptop .....	27

### TÓM TẮT

Trực quan hóa dữ liệu bằng AWS là một quá trình sử dụng công cụ phân tích dữ liệu của AWS để biểu diễn thông tin dưới dạng đồ thị, biểu đồ hoặc hình ảnh, giúp người dùng dễ dàng và trực quan hóa dữ liệu. AWS cung cấp nhiều công cụ và chức năng để hỗ trợ quá trình này.

Xác định rõ mục tiêu của việc trực quan hóa dữ liệu. Điều này sẽ giúp xác định loại dữ liệu cần thu thập và cách hiển thị thông tin một cách phù hợp.

Thu thập dữ liệu từ nguồn đáng tin cậy như cơ sở dữ liệu, bảng tính, hoặc khảo sát để đảm bảo chất lượng và độ chính xác của dữ liệu.

Tiến hành xử lý dữ liệu, bao gồm làm sạch, loại bỏ lỗi và chuẩn hóa dữ liệu để đảm bảo tính nhất quán và đồng đều.

Chọn loại biểu đồ hoặc đồ thị phù hợp với mục tiêu hiển thị dữ liệu. Sử dụng biểu đồ phân phối, biểu đồ tương quan, biểu đồ cây quyết định, biểu đồ mạng, hoặc biểu đồ thời gian tùy thuộc vào loại thông tin bạn muốn truyền đạt.

Sử dụng công cụ trực quan hóa dữ liệu của AWS hoặc tạo thủ công bằng phần mềm đồ họa để tạo biểu đồ hoặc đồ thị. Đảm bảo rằng chúng rõ ràng và dễ hiểu.

Tận dụng màu sắc và biểu tượng một cách có chủ ý. Sử dụng chúng để làm cho dữ liệu trở nên trực quan hóa và dễ nhìn.

Bổ sung nhãn và chú thích rõ ràng để giúp người xem hiểu thông tin. Chú ý đến việc cung cấp thông tin cần thiết để giảm hiểu lầm.

Trước khi tạo biểu đồ hoặc đồ thị, kiểm tra kỹ lưỡng độ chính xác của dữ liệu để tránh hiểu lầm và kết quả trực quan không chính xác.



## MỞ ĐẦU

### 1. Lý do chọn đề tài

- Trong thời đại số hóa ngày nay, dữ liệu đóng vai trò vô cùng quan trọng trong mọi lĩnh vực của đời sống, từ kinh doanh, sản xuất, dịch vụ đến nghiên cứu khoa học. Việc thu thập, phân tích và xử lý dữ liệu là một công việc phức tạp và đòi hỏi nhiều kỹ thuật cao.
- Trực quan hóa dữ liệu đóng vai trò quan trọng trong phân tích dữ liệu, chuyển đổi thông tin thành các biểu đồ, đồ thị và hình ảnh dễ hiểu. Điều này sẽ giúp ta nhận thức sâu sắc hơn về dữ liệu giúp hỗ trợ quyết định hiệu quả hơn.
- Công cụ AWS là nền tảng điện đám mây toàn diện cho việc phân tích dữ liệu, bao gồm các dịch vụ lưu trữ dữ liệu, kho dữ liệu, máy học và phân tích dữ liệu lớn. AWS cung cấp các công cụ trực quan hóa dữ liệu mạnh mẽ như Amazon QuickSight.

### 2. Mục đích

- Hiểu rõ hơn về tính chất và cấu trúc của dữ liệu, phát hiện ra các mối quan hệ giữa các thuộc tính và dữ liệu.
- Phát hiện và giải quyết các vấn đề liên quan đến dữ liệu gồm các giá trị bất thường, các giá trị thiếu, các biến động và nhiễu dữ liệu.
- Truyền tải thông tin một cách trực quan và dễ hiểu hơn giúp người xem dễ dàng tiếp cận và hiểu rõ hơn về dữ liệu.
- Đưa ra quyết định dựa trên dữ liệu, hiểu rõ hơn về tình hình và có thể đưa ra các quyết định chính xác hơn.
- Khám phá dữ liệu, phát hiện ra các mẫu, xu hướng và quan hệ giữa các biến trong dữ liệu, giúp ta có thể đưa ra các phân tích và dự đoán chính xác hơn.

### 3. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu là các dữ liệu ở dạng số, văn bản hoặc hình ảnh. Dữ liệu này có thể được thu thập từ nhiều nguồn khác nhau, chẳng hạn như bảng tính, cơ sở dữ liệu, mạng xã hội hoặc khảo sát.

- Nghiên cứu lý thuyết: Khái niệm, mục đích, lợi ích và ứng các loại hình và phương pháp trực quan hóa dữ liệu.
- Nghiên cứu và sử dụng công cụ trực quan hóa dữ liệu của AWS, bao gồm: Amazon QuickSight, Amazon Sage Maker, Amazon Athena, Amazon EMR...
- Áp dụng các công cụ AWS và trực quan hóa dữ liệu để phân tích dữ liệu thực tế từ một lĩnh vực cụ thể.

## CHƯƠNG 1: TỔNG QUAN

### 1.1 Quá trình Trực Quan hóa Dữ Liệu trên AWS:

- Trực quan hóa dữ liệu không chỉ là việc hiển thị số liệu mà còn là cách để tạo ra sự hiểu biết sâu rộng về dữ liệu. Các công cụ phân tích dữ liệu của AWS cung cấp một nền tảng mạnh mẽ để biểu diễn dữ liệu dưới dạng đồ thị, biểu đồ và hình ảnh, giúp người dùng hiểu rõ hơn về cấu trúc và xu hướng của dữ liệu, nhanh chóng xác định các mô hình và sự tương quan và đưa ra quyết định dựa trên thông tin một cách nhanh chóng và chính xác.

### 1.2. Đặc Điểm Của AWS Trong Trực Quan hóa Dữ Liệu:

- Linh hoạt và Mở Rộng: AWS cung cấp các dịch vụ đám mây linh hoạt, giúp người dùng mở rộng tài nguyên theo nhu cầu và thích ứng với quy mô dữ liệu.
- Bảo Mật và Tuân Thủ: AWS chú trọng đến bảo mật dữ liệu và tuân thủ các quy tắc an ninh, đảm bảo rằng thông tin được trực quan hóa luôn an toàn và tuân thủ các chuẩn quy định.
- Tích Hợp Công Cụ Trực Quan Hóa: AWS tích hợp nhiều công cụ trực quan hóa như Amazon QuickSight, giúp người dùng tạo và tùy chỉnh các biểu đồ một cách thuận tiện.

### 1.3. Mục tiêu nghiên cứu và cách giải Quyết:

- Nghiên cứu và hiểu rõ về các công cụ phân tích dữ liệu của AWS.
- Tìm hiểu cách tích hợp và sử dụng các tính năng trực quan hóa dữ liệu trên AWS.
- Áp dụng các kỹ thuật trực quan hóa để hiểu sâu về dữ liệu và đưa ra những thông tin quan trọng.

## **CHƯƠNG 2: NGHIÊN CỨU LÝ THUYẾT**

### **2.1 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 biểu đồ, đồ thị 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.
- **Trực quan hóa dữ liệu mang lại nhiều lợi ích:**
  - Dễ dàng hiểu và phân tích: Trực quan hóa dữ liệu giúp chúng ta dễ dàng nhìn thấy mô hình, mối quan hệ, sự phân bố, xu hướng và biểu hiện của dữ liệu, từ đó giúp chúng ta hiểu dữ liệu một cách nhanh chóng và chính xác hơn.
  - Tăng tính tương tác: Việc sử dụng các đồ thị, biểu đồ, hình ảnh trực quan giúp người dùng tương tác với dữ liệu và phân tích dữ liệu một cách linh hoạt hơn.
  - Tối ưu hóa quyết định: Dữ liệu được trực quan hóa sẽ giúp người sử dụng dễ dàng xác định và đưa ra quyết định một cách nhanh chóng và chính xác hơn.
  - Hiển thị mối quan hệ giữa các biến: Trực quan hóa dữ liệu cũng giúp người dùng dễ dàng nhận thấy mối quan hệ giữa các biến trong dữ liệu, từ đó giúp họ tìm ra được các thông tin quan trọng.
  - Giảm thời gian phân tích dữ liệu: Sử dụng các công cụ trực quan hóa dữ liệu giúp người dùng tiết kiệm thời gian phân tích và đưa ra kết luận chính xác hơn.
  - Tăng tính chính xác và độ tin cậy: Trực quan hóa dữ liệu giúp người dùng tránh được các sai sót do đọc và phân tích dữ liệu bằng cách đơn giản hóa quá trình này và giúp dữ liệu được hiển thị một cách chính xác và đáng tin cậy hơn

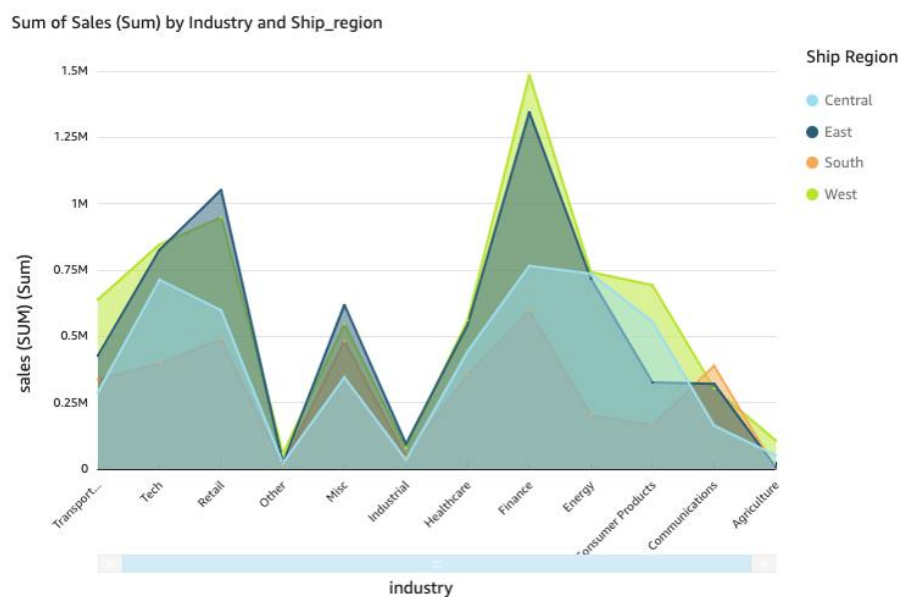
#### **2.1.2 Các thành phần trực quan hóa dữ liệu**

- **Trục dữ liệu:** là các thành phần trong biểu đồ nó thể hiện giá trị của dữ liệu. Trục dữ liệu có thể là trục x, trục y hoặc cả hai. Trục dữ liệu được chia thành các đơn vị giúp người dùng dễ dàng đọc và hiểu về dữ liệu.
- **Biểu đồ:** là cách thể hiện dữ liệu dưới dạng hình ảnh giúp người dùng dễ dàng hiểu và phân tích dữ liệu.

- Dữ liệu: Là các giá trị được đặt trên trục dữ liệu và được trực quan hóa trong biểu đồ.
- Nhãn: Là các nhãn tiêu đề hoặc chú thích giúp giải thích và định nghĩa cho các thành phần của trực quan hóa dữ liệu, giúp người dùng hiểu rõ hơn về thông tin được trình bày.
- Công cụ phân tích: Là các công cụ hỗ trợ người dùng phân tích dữ liệu trên trực quan hóa dữ liệu, giúp tìm ra mối quan hệ, xu hướng và thông tin quan trọng từ dữ liệu.
- Bố cục: Là cách sắp xếp các thành phần của trực quan hóa dữ liệu, giúp cho người dùng dễ dàng nhận thấy các yếu tố quan trọng của dữ liệu
- Màu sắc: Là yếu tố quan trọng trong việc trực quan hóa dữ liệu, giúp tăng tính thẩm mỹ và truyền tải thông tin một cách rõ ràng và phân biệt các giá trị và nhóm dữ liệu.

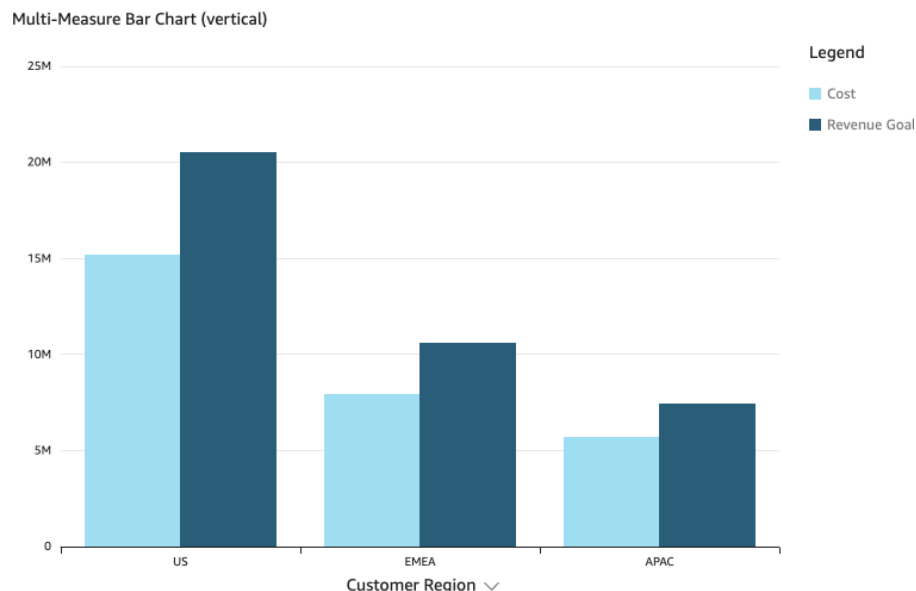
### 2.1.3 Các loại biểu đồ

- **Line chart:** được sử dụng để biểu diễn sự thay đổi của một dữ liệu theo thời gian hoặc một biến độc lập khác. Thông qua việc kết nối các điểm dữ liệu bằng đường, biểu đồ đường giúp người quan sát nhận thức mối liên quan và xu hướng của dữ liệu một cách rõ ràng.



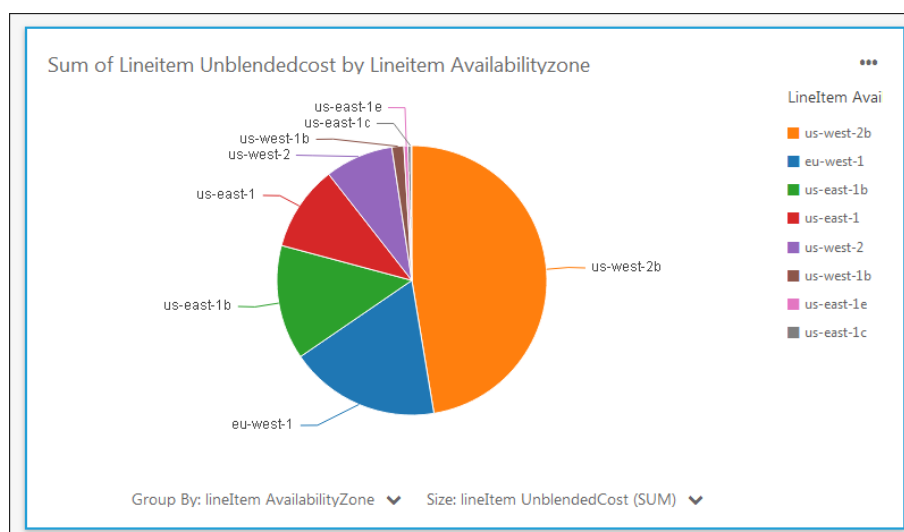
Hình 2. 1 Line chart

- **Bar chart:** Biểu đồ cột được sử dụng để so sánh giá trị của các biến độc lập khác nhau. Cấu trúc của biểu đồ cột thường bao gồm các cột dọc đứng, mỗi cột thể hiện giá trị của một biến độc lập cụ thể. Độ dài của cột thường tương ứng với giá trị của biến đó, tạo ra một biểu đồ dễ hiểu và trực quan.



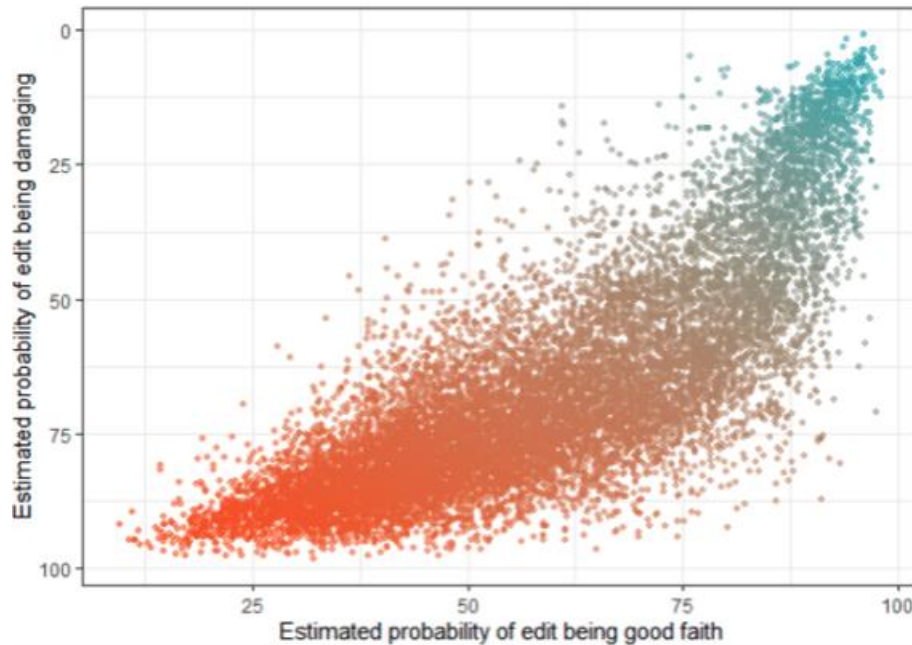
Hình 2. 2 Bar chart

- **Pie chart:** Biểu đồ hình bánh giúp thể hiện sự phân phối của dữ liệu một cách rõ ràng và dễ hiểu. Mỗi phần trăm được thể hiện qua diện tích của phần tương ứng trên đồ tròn. Các mảng màu sắc hoặc các ký hiệu khác nhau thường được sử dụng để làm nổi bật và phân biệt giữa các phần khác nhau



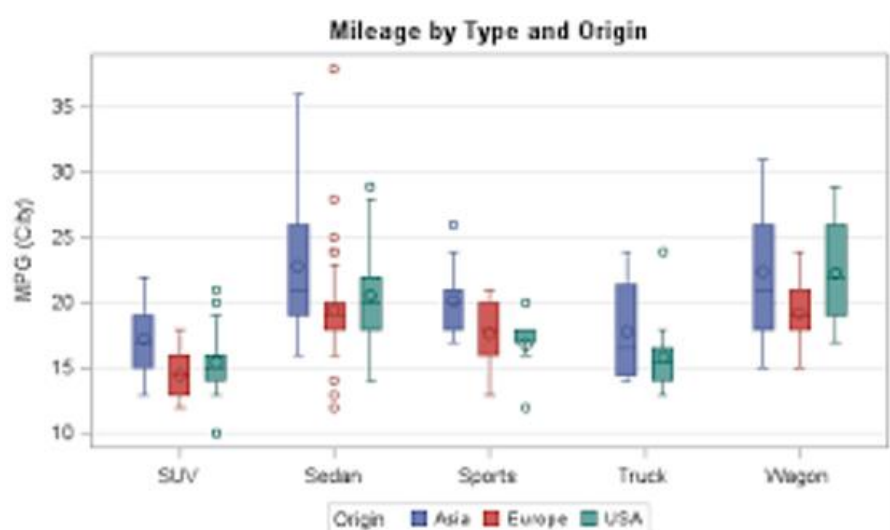
Hình 2. 3 Pie chart

- **Scatter plot:** Biểu đồ phân tán, hay Scatter plot được sử dụng để thể hiện mối tương quan giữa hai biến độc lập. Trong biểu đồ này, mỗi điểm dữ liệu được đặc trưng bởi một cặp giá trị tương ứng trên trục x và trục y. Sự phân tán của các điểm trên biểu đồ phản ánh mối quan hệ giữa hai biến, giúp nhận diện xu hướng hoặc mô hình tương quan.



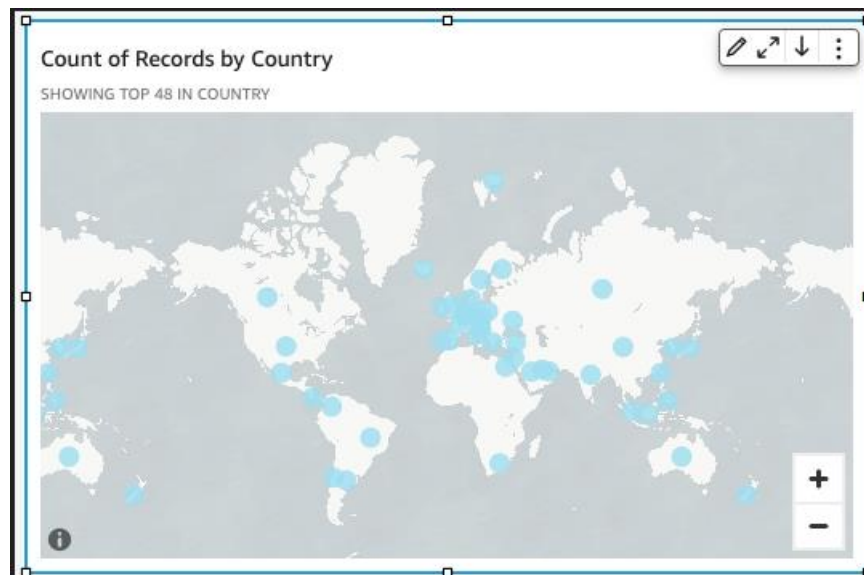
Hình 2. 4 Scatter plot

- **Box plot:** Biểu đồ hộp được sử dụng để biểu diễn phân bố của dữ liệu.



Hình 2. 5 Box plot

- **Map:** được sử dụng để biểu diễn dữ liệu dưới dạng không gian và địa lý. Bằng cách ánh xạ thông tin lên một bề mặt phẳng, bản đồ giúp chúng ta thấy rõ mối quan hệ và phân bố của dữ liệu trên bản đồ địa lý.



Hình 2. 6 Map

### 2.1.4 Các bước trực quan hóa dữ liệu

#### ❖ Thu thập dữ liệu từ các nguồn đa dạng:

- Quá trình này bắt đầu bằng việc xác định các nguồn dữ liệu chính từ cơ sở dữ liệu, tệp dữ liệu, hoặc thiết bị đo lường. Điều này yêu cầu sự chú ý đến tính đầy đủ và độ tin cậy của dữ liệu thu thập.

#### ❖ Khám phá dữ liệu:

- Bước này tập trung vào việc sử dụng các phương pháp thống kê và khám phá dữ liệu để hiểu rõ hơn về đặc tính của dữ liệu. Phân tích biểu đồ, phân phối dữ liệu, và xác định mối quan hệ giữa các biến giúp xây dựng cơ sở hiểu biết sâu sắc. Các giả thuyết ban đầu cũng được đưa ra để hướng dẫn quá trình khám phá một cách có mục đích.

#### ❖ Tiền xử lý dữ liệu:

- Bước tiền xử lý dữ liệu trở nên quan trọng để chuẩn bị dữ liệu cho việc trực quan hóa. Làm sạch dữ liệu bằng cách lọc nhiễu, điền giá trị bị khuyết một cách logic, chuyển đổi định dạng dữ liệu, và chuẩn hóa giúp đảm bảo tính nhất quán và độ tin cậy.



### ❖ **Lựa chọn phương pháp trực quan hóa:**

- Quyết định phương pháp trực quan hóa phù hợp chọn loại biểu đồ phản ánh tính chất của dữ liệu và mục đích trực quan hóa

### ❖ **Xây dựng trực quan hóa:**

- Tạo biểu đồ, biểu đồ tần số, bản đồ, hình ảnh và video để truyền đạt thông tin một cách hấp dẫn và dễ hiểu. Sự sáng tạo trong cách trình bày thông tin làm cho trực quan hóa trở nên sinh động và thu hút sự chú ý.

### ❖ **Hiệu chỉnh và tinh chỉnh:**

- Sau khi xây dựng, quá trình hiệu chỉnh và tinh chỉnh là quan trọng để đảm bảo rằng trực quan hóa thể hiện đầy đủ thông tin và dễ hiểu cho khán giả. Điều chỉnh màu sắc, kích thước và định dạng là quan trọng để tối ưu hóa trực quan hóa. Phản hồi từ người xem cũng đóng một vai trò quan trọng trong quá trình này.

### ❖ **Xuất bản và chia sẻ:**

- Sau khi hoàn thiện, trực quan hóa được xuất bản và chia sẻ với khán giả mục tiêu qua nhiều phương tiện truyền thông như trang web, báo chí, mạng xã hội hoặc báo cáo đảm bảo rằng thông điệp được truyền đạt một cách hiệu quả và đến đúng đối tượng.

## 2.1.5 Ứng dụng

- **Kinh doanh:** Trực quan hóa dữ liệu được sử dụng để phân tích dữ liệu kinh doanh, chẳng hạn như dữ liệu về khách hàng, dữ liệu tài chính, và dữ liệu bán hàng. Việc phân tích này có thể giúp các doanh nghiệp đưa ra các quyết định kinh doanh hiệu quả và sáng suốt hơn.
- **Sản xuất:** Trực quan hóa dữ liệu được sử dụng để phân tích dữ liệu sản xuất, chẳng hạn như dữ liệu về hiệu suất, dữ liệu an toàn và dữ liệu chất lượng. Việc phân tích này có thể giúp các nhà sản xuất cải thiện chất lượng sản phẩm hiệu quả trong sản xuất và an toàn lao động.
- **Dịch vụ:** Trực quan hóa dữ liệu được sử dụng để phân tích dữ liệu dịch vụ chẳng hạn như dữ liệu khách hàng, dữ liệu nhân viên, dữ liệu hoạt động có thể giúp các tổ chức dịch vụ cải thiện trải nghiệm khách hàng và hiệu quả trong hoạt động.

- Nghiên cứu khoa học: Trực quan hóa dữ liệu được sử dụng để phân tích dữ liệu nghiên cứu khoa học, chẳng hạn như dữ liệu từ thí nghiệm, dữ liệu quan sát và dữ liệu mô phỏng.

## 2.2 Amazon Web Service

### 2.2.1 Khái niệm

- Công cụ AWS là bản viết tắt của "Amazon Web Services", đây là một dịch vụ đám mây được cung cấp bởi Amazon.com. AWS cung cấp một loạt các dịch vụ máy chủ, lưu trữ, cơ sở dữ liệu, mạng, công cụ phân tích dữ liệu, trí tuệ nhân tạo (AI), máy học và nhiều dịch vụ khác để giúp doanh nghiệp và cá nhân nhân viên xây dựng và quản lý ứng dụng trực tuyến, lưu trữ dữ liệu và phát triển dự án trên môi trường đám mây.

### 2.2.2 Amazon QuickSight

- QuickSight, là dịch vụ trực quan hóa dữ liệu hàng đầu của AWS, công cụ cho phép dễ dàng kết nối truy vấn dữ liệu đa dạng và kết nối với nhiều nguồn dữ liệu khác nhau, QuickSight giúp người dùng truy cập nhanh nhanh chóng vào dữ liệu lưu trữ trên các công cụ Amazon S3, Amazon RDS, Amazon Redshift và nhiều nguồn dữ liệu khác.
- QuickSight cho phép người dùng truy cập dữ liệu dễ dàng và giúp người dùng tạo biểu đồ và bảng điều khiển trực quan hóa dữ liệu. Với đa dạng các loại biểu đồ và đồ thị khác nhau giúp người dùng có thể linh hoạt hiển thị và hiểu dữ liệu theo cách phù hợp với mục đích của mình.
- Sự tích hợp thông minh của QuickSight cho phép người dùng có quyền truy cập vào dữ liệu từ nhiều nguồn khác nhau giúp họ phân tích dữ liệu một cách toàn diện và đưa ra quyết định sáng suốt hơn trong quá trình phân tích dữ liệu.

### 2.2.3 Amazon SageMaker

- Amazon SageMaker là một dịch vụ máy học toàn diện của AWS, được thiết kế để cung cấp đa dạng khả năng xử lý và phân tích dữ liệu. Nó tích hợp nhiều thuật toán học máy phổ biến để hỗ trợ phân tích và hiểu dữ liệu một cách hiệu

quả. Điều này không chỉ biến dịch vụ thành một môi trường đào tạo mô hình mà còn tạo ra một hệ sinh thái máy học toàn diện.

- Amazon SageMaker mang đến nhiều tính năng hữu ích, bao gồm khả năng hiệu quả xử lý dữ liệu từ việc thu thập đến làm sạch và chuẩn hóa cho quá trình đào tạo mô hình. Dịch vụ này tích hợp nhiều thuật toán học máy phổ biến, từ phân loại đến dự đoán và gom cụm, giúp người dùng chọn lựa phương pháp phù hợp với mục tiêu của họ.
- AWS SageMaker còn tích hợp với các công cụ trực quan hóa dữ liệu, giúp người dùng dễ dàng hiểu rõ kết quả của mô hình thông qua biểu đồ và biểu đồ phân phối. Người dùng cũng có khả năng triển khai mô hình một cách thuận lợi và tích hợp nó vào ứng dụng và quy trình công việc của mình.

#### **2.2.4 Amazon Athena**

- Amazon Athena là một dịch vụ truy vấn dữ liệu của AWS, được thiết kế để cho phép người dùng truy vấn và phân tích dữ liệu được lưu trữ trong Amazon S3 bằng ngôn ngữ truy vấn SQL. Dịch vụ này giúp đơn giản hóa quá trình truy xuất dữ liệu và tạo thuận tiện cho việc thực hiện các phân tích SQL trực tiếp trên dữ liệu lưu trữ trong kho S3.
- Athena cho phép người dùng truy vấn dữ liệu trong Amazon S3 mà không cần phải di chuyển hoặc sao chép dữ liệu đến một hệ thống lưu trữ khác.
- Sử dụng ngôn ngữ truy vấn SQL phổ biến, Athena tạo điều kiện cho việc truy vấn và xử lý dữ liệu một cách thuận lợi, giảm độ phức tạp cho người dùng.
- Dịch vụ cung cấp khả năng trực quan hóa kết quả truy vấn thông qua biểu đồ và bảng điều khiển, giúp người dùng hiểu rõ hơn về dữ liệu và kết quả của các truy vấn.
- Athena là một dịch vụ có khả năng mở rộng linh hoạt, có thể xử lý các tập dữ liệu lớn và đồng thời hỗ trợ nhiều loại định dạng dữ liệu.

#### **2.2.5. Amazon Kinesis**

- Amazon Kinesis là một dịch vụ stream dữ liệu của AWS, được thiết kế để cung cấp khả năng thu thập, xử lý và phân tích dữ liệu thời gian thực. Dịch vụ này

giúp người dùng xử lý lưu lượng dữ liệu lớn và đồng thời hỗ trợ các tính năng trực quan hóa dữ liệu để hiển thị dữ liệu đang được stream trực tiếp.

- Kinesis cho phép thu thập và xử lý dữ liệu thời gian thực, giúp người dùng làm việc với dữ liệu đang được tạo ra ngay lập tức và tối ưu hóa để xử lý lưu lượng dữ liệu lớn, giúp người dùng quản lý và định hình dữ liệu từ nguồn đến đích một cách hiệu quả.
- Kinesis cung cấp khả năng xử lý dữ liệu ngay khi nó được tạo ra, giúp giảm độ trễ và tăng tính khả dụng của dữ liệu.
- Dịch vụ này tích hợp với các tính năng trực quan hóa dữ liệu, giúp người dùng theo dõi và hiển thị thông tin từ dữ liệu đang được stream một cách trực quan.
- Dịch vụ này cung cấp các tính năng bảo mật như mã hóa dữ liệu và quản lý quyền truy cập để đảm bảo an toàn cho dữ liệu.

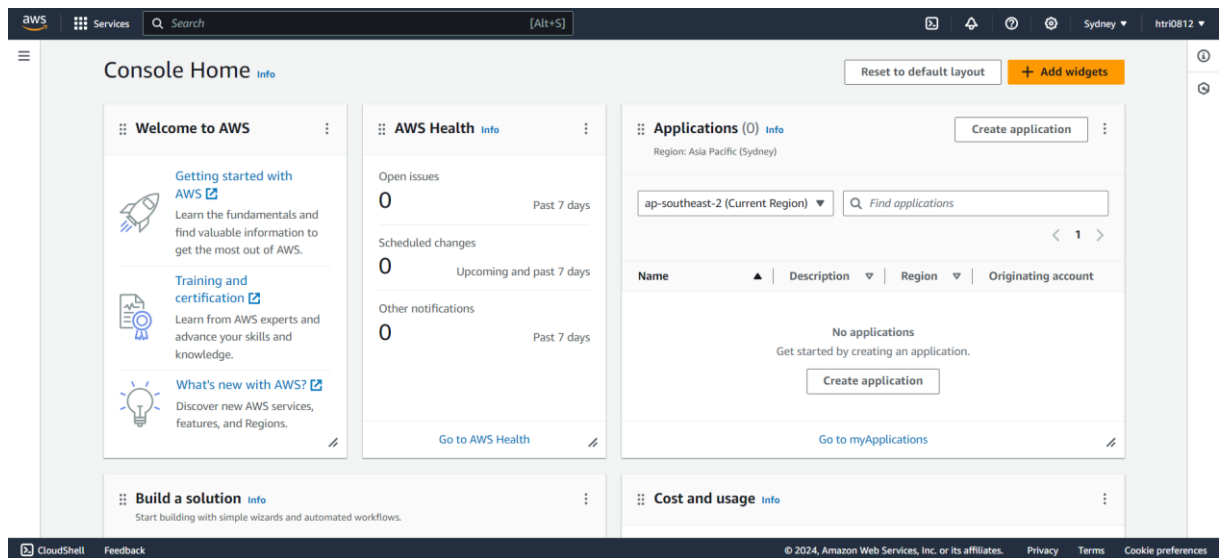
#### **2.2.6 Amazon EMR**

- Amazon EMR là một dịch vụ xử lý phân tán được thiết kế để đơn giản hóa quá trình xử lý và phân tích dữ liệu lớn bằng cách sử dụng các công cụ mã nguồn mở như Apache Spark, Apache Hive, Apache Hadoop, và nhiều công cụ khác. Dịch vụ này giúp người dùng mở rộng linh hoạt và hiệu quả các tài nguyên tính toán để xử lý khối lượng lớn dữ liệu một cách hiệu quả.
- EMR cho phép xử lý phân tán dữ liệu lớn thông qua việc chia nhỏ tác vụ và phân phối chúng đến các node xử lý khác nhau, giúp tăng tốc quá trình xử lý và giảm thời gian thực hiện.
- Người dùng có thể linh hoạt tùy chỉnh kích thước của cụm xử lý để đáp ứng nhu cầu cụ thể và tự động mở rộng tài nguyên tính toán theo yêu cầu.
- EMR tích hợp với các công cụ trực quan hóa dữ liệu như Amazon QuickSight, giúp người dùng trực quan hóa kết quả của quá trình phân tích dữ liệu một cách dễ dàng và sinh động.
- Cung cấp các tính năng bảo mật như mã hóa dữ liệu, quản lý quyền truy cập, và tích hợp quản lý kiểm soát quyền truy cập.

## CHƯƠNG 3: HIỆN THỰC HÓA NGHIÊN CỨU

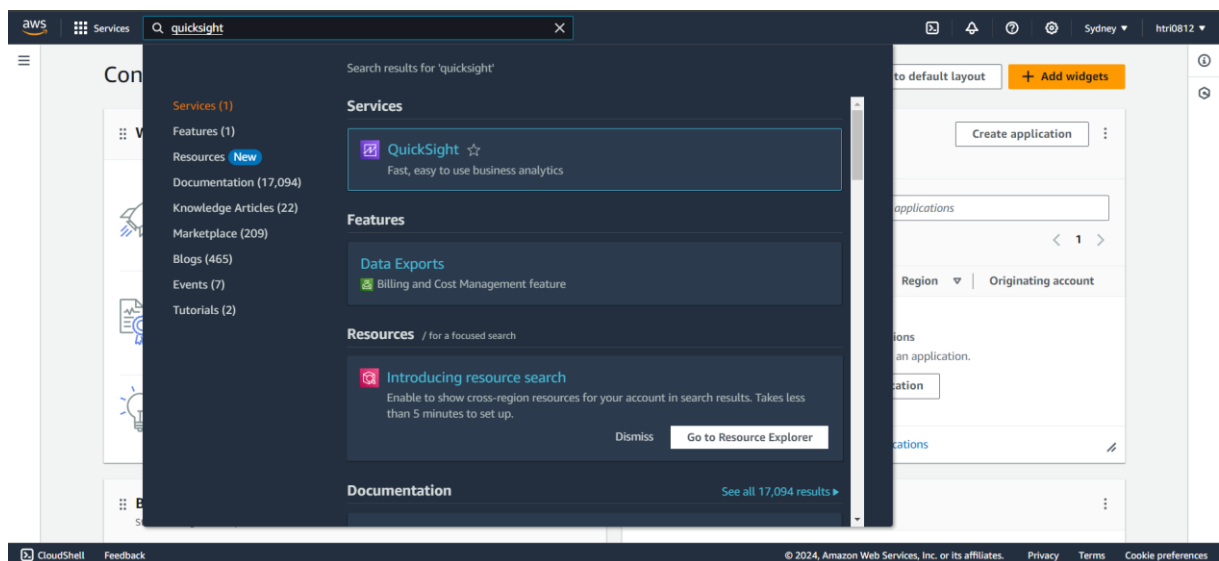
### 3.1 Áp dụng trực quan hóa dữ liệu trên AWS

#### 3.1.1 Giao diện chính của công cụ



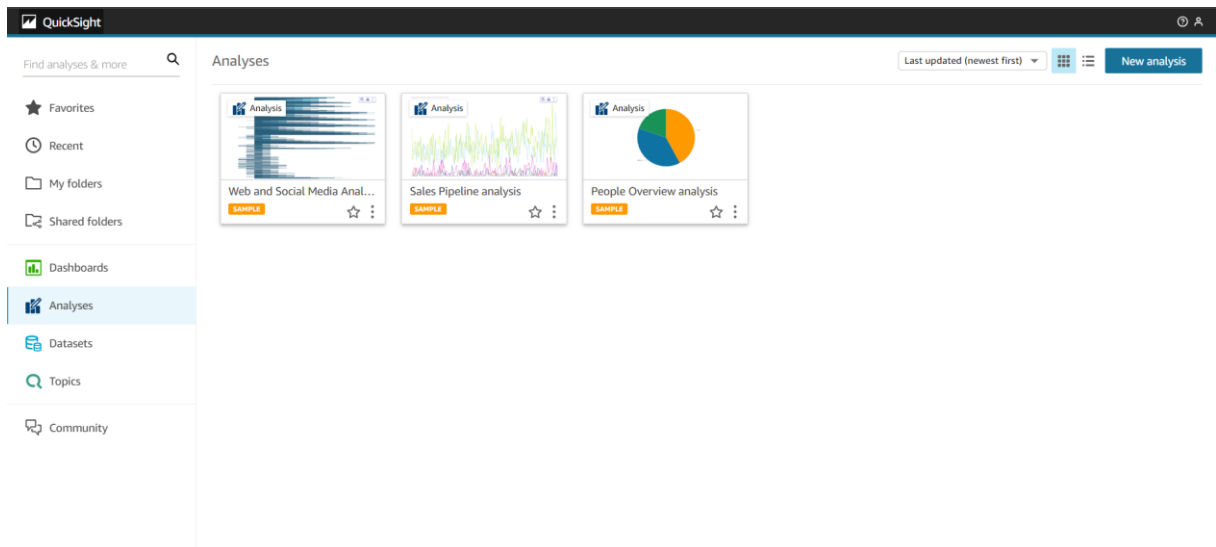
Hình 3. 1 Giao diện chính của công cụ AWS

- Để truy cập vào công cụ AWS QuickSight ta điều hướng đến bảng điều khiển tìm và chọn dịch vụ “QuickSight”.



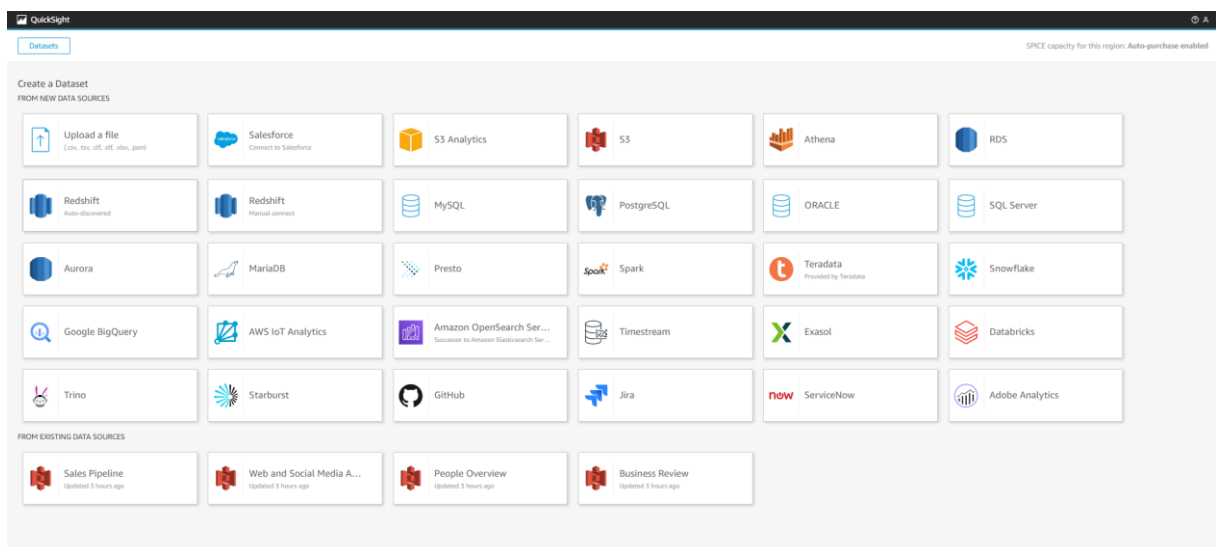
Hình 3. 2 Giao diện tìm và truy cập dịch vụ AWS QuickSight

- Chọn dataset hiện có hoặc tạo dataset từ dữ liệu bạn muốn trực quan hóa. Có thể là các dữ liệu trong Amazon S3, Redshift, RDS, hoặc từ nguồn dữ liệu khác.



Hình 3. 3 Giao diện tạo dataset AWS QuickSight

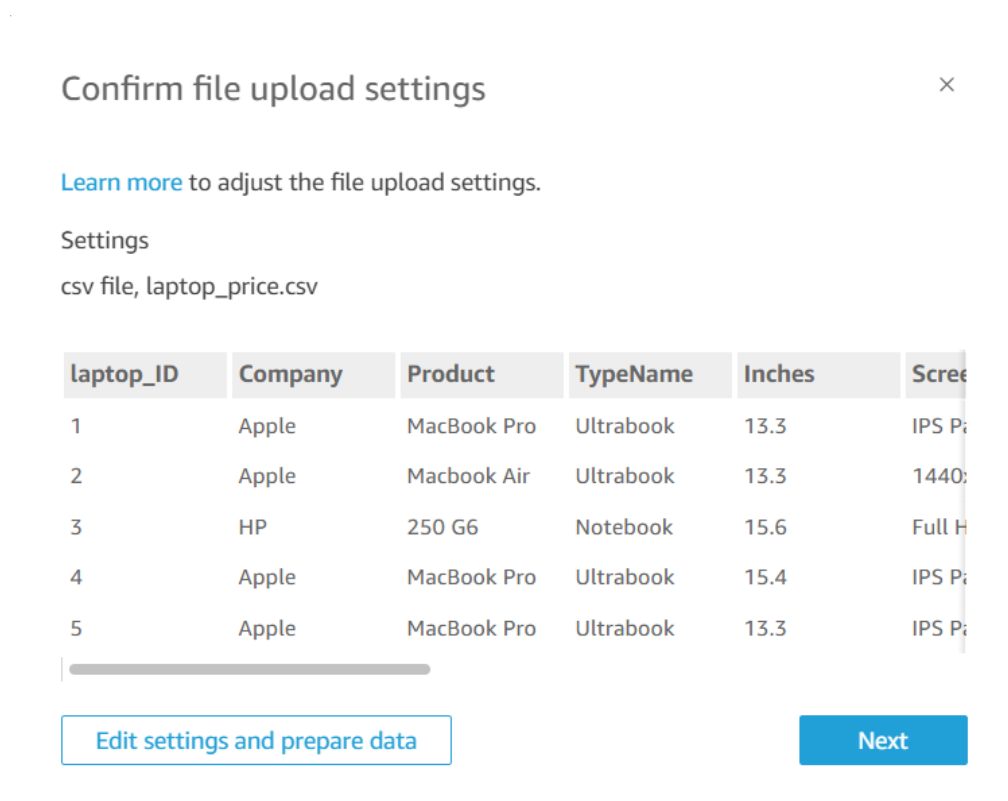
- Ta chọn biểu tượng “New analysis” để tạo một bảng phân tích mới. Khi nhấp vào đây sẽ yêu cầu bạn chọn bất kỳ tập dữ liệu nào. Ta cũng có thể tạo một tập dữ liệu mới khi chọn “New dataset”



Hình 3. 4 Giao diện tìm và truy cập dịch vụ AWS QuickSight

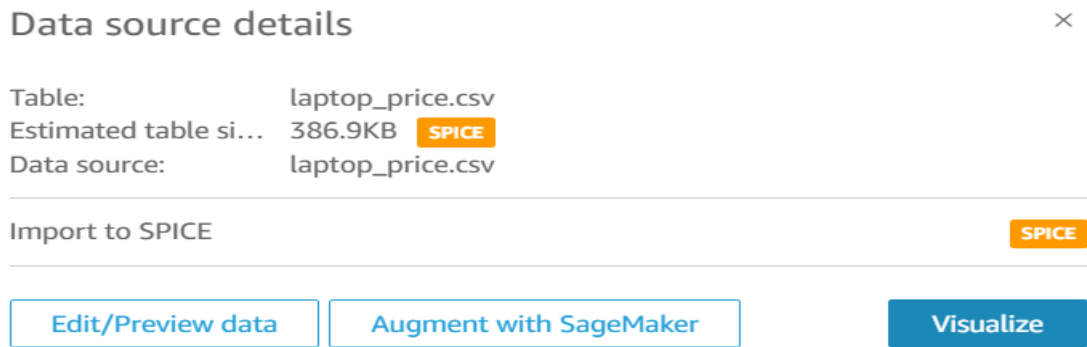
- Quick Sight hỗ trợ trực quan hóa dữ liệu từ nhiều loại tập dữ liệu khác nhau bao gồm:

- **File csv** (comma-separated values): Đây là loại file dữ liệu phổ biến nhất, được định dạng bằng các dấu phẩy để phân tách các giá trị dữ liệu.
  - **File json** (JavaScript Object Notation): Đây là loại file dữ liệu được định dạng bằng JSON.
  - **File xlsx** (Excel): Đây là loại file dữ liệu được tạo bởi Microsoft Excel.
  - **File txt** (text): Đây là loại file văn bản, không có định dạng đặc biệt.
  - **Các cơ sở dữ liệu**, chẳng hạn như MySQL, PostgreSQL, v.v.
  - **Các API**, chẳng hạn như API của Google Analytics, v.v.
  - **Các dịch vụ web**, chẳng hạn như dịch vụ web của Google Maps, v.v.
- Chọn "Upload a file" và tải lên tệp CSV hoặc Excel chọn Next để tiếp tục.
  - Ta có: Chuẩn bị tập dữ liệu về Laptop Price để chuẩn bị cho phân tích.



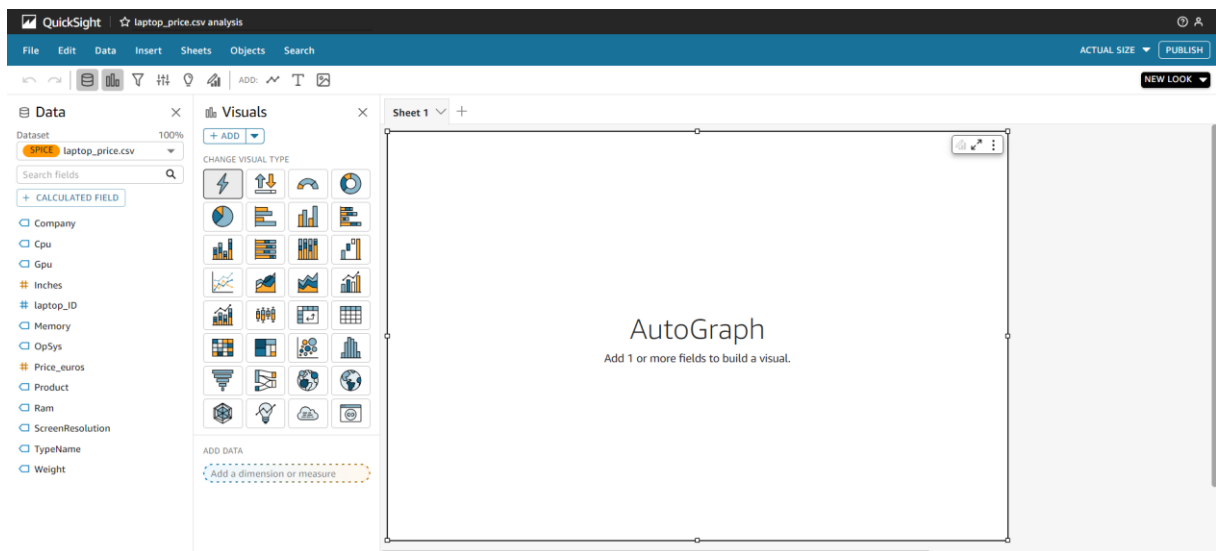
Hình 3. 5 Giao diện tập dữ liệu được chọn AWS QuickSight

- Khi tập dữ liệu đã được chọn, ta chọn tùy chọn "Visualize" để bắt đầu quá trình trực quan hóa.



Hình 3. 6 Giao diện tìm và truy cập dịch vụ AWS QuickSight

- Trong giao diện QuickSight, tìm đến và mở "Field List". Đây là nơi hiển thị danh sách các trường trong tập dữ liệu của bạn.

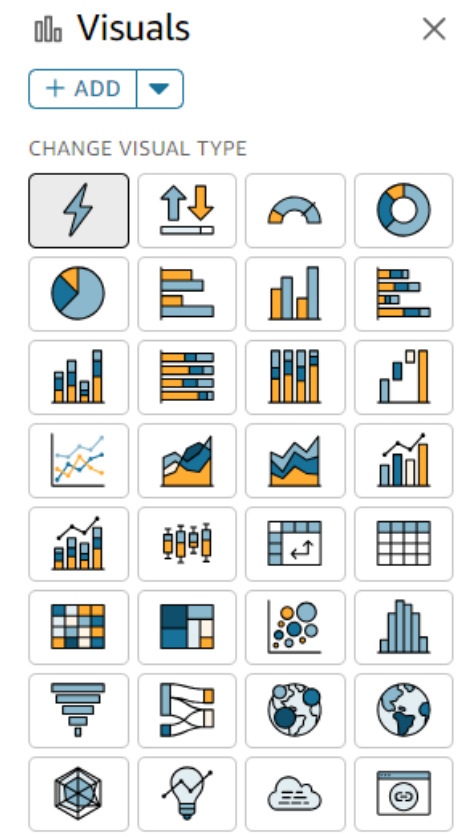


Hình 3. 7 Giao diện tìm và truy cập dịch vụ AWS QuickSight

### 3.1.2 Tạo biểu đồ trực quan hóa

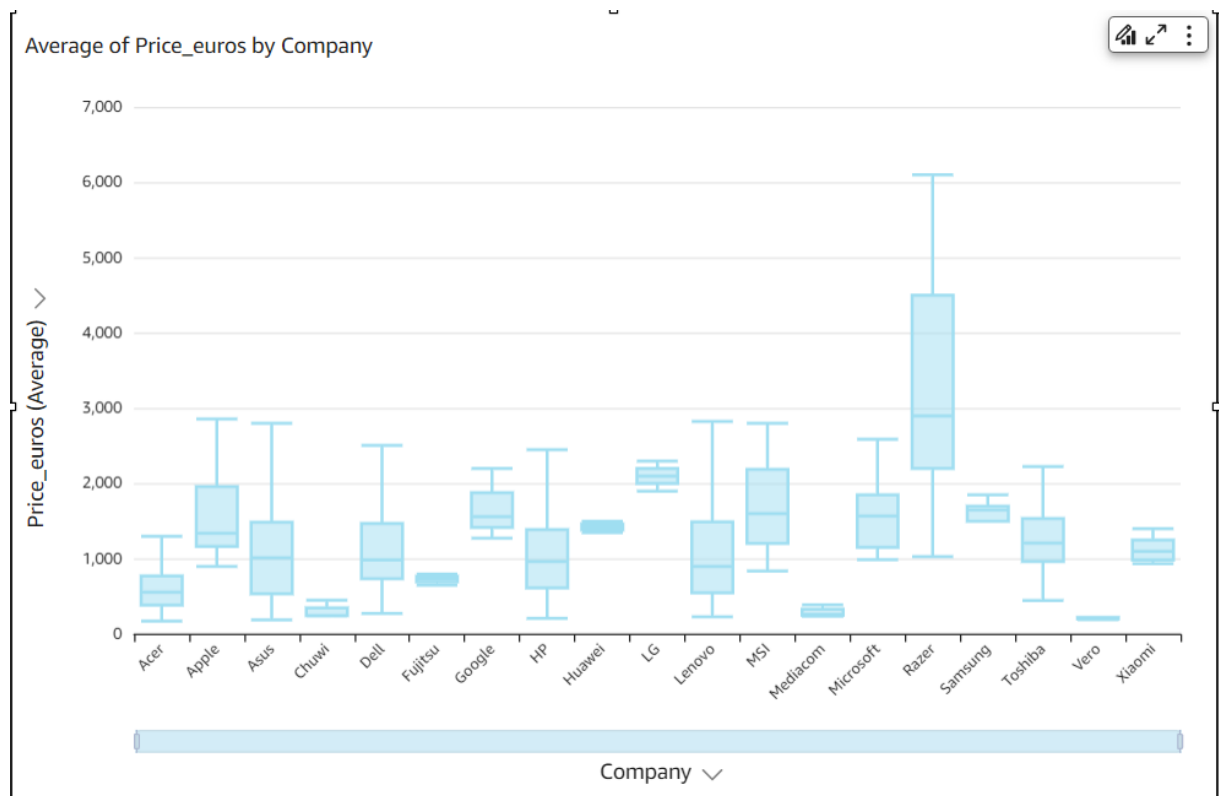
- **Trên trang phân tích:** chọn **Add** trên thanh ứng dụng. Một hình ảnh trống, mới được tạo và AutoGraph được chọn theo mặc định.



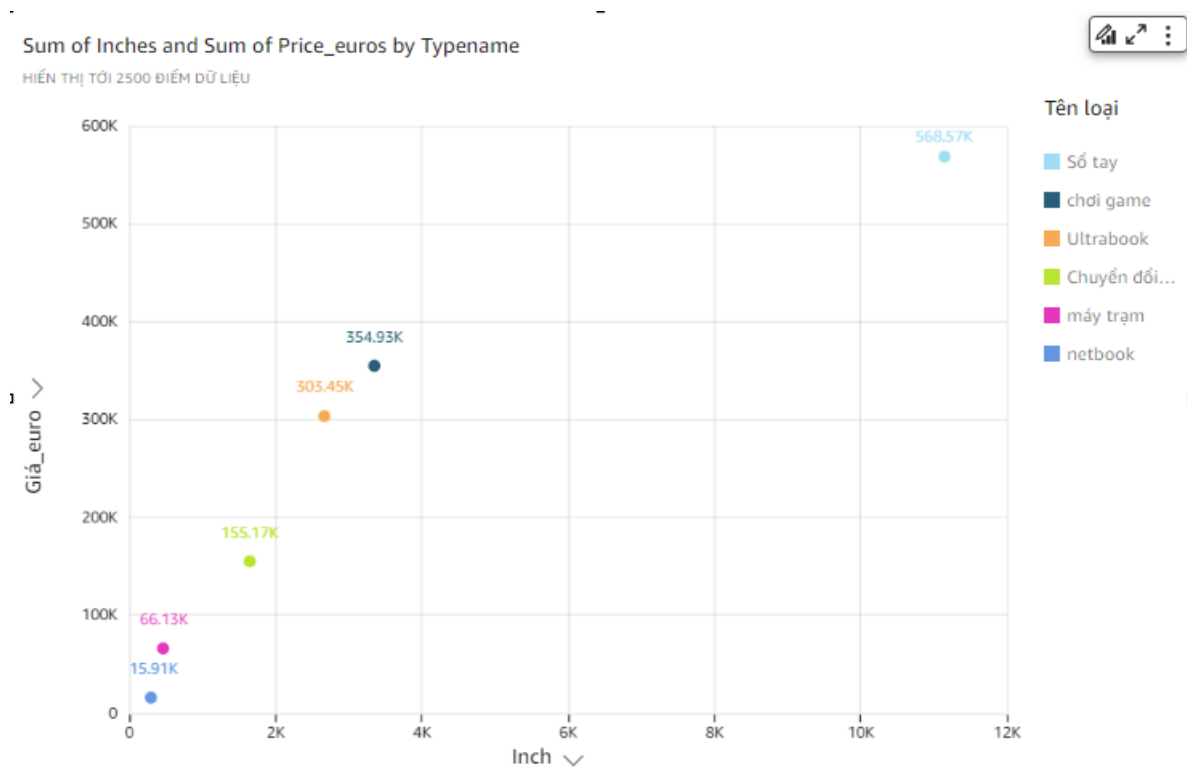


Hình 3. 8 Chức năng Visuals

- Chọn loại biểu đồ trực quan hóa: Có thể thay đổi các loại hình ảnh biểu đồ phù hợp để phân tích dữ liệu.
- Chọn dữ liệu cho biểu đồ: Chọn trường dữ liệu từ dataset của bạn để ánh xạ vào trục x và trục y và nhóm của biểu đồ. Các lựa chọn này phụ thuộc vào loại biểu đồ bạn đang sử dụng.
- Sử dụng biểu đồ hộp để thể hiện phân phối giá của laptop theo từng nhà sản xuất.

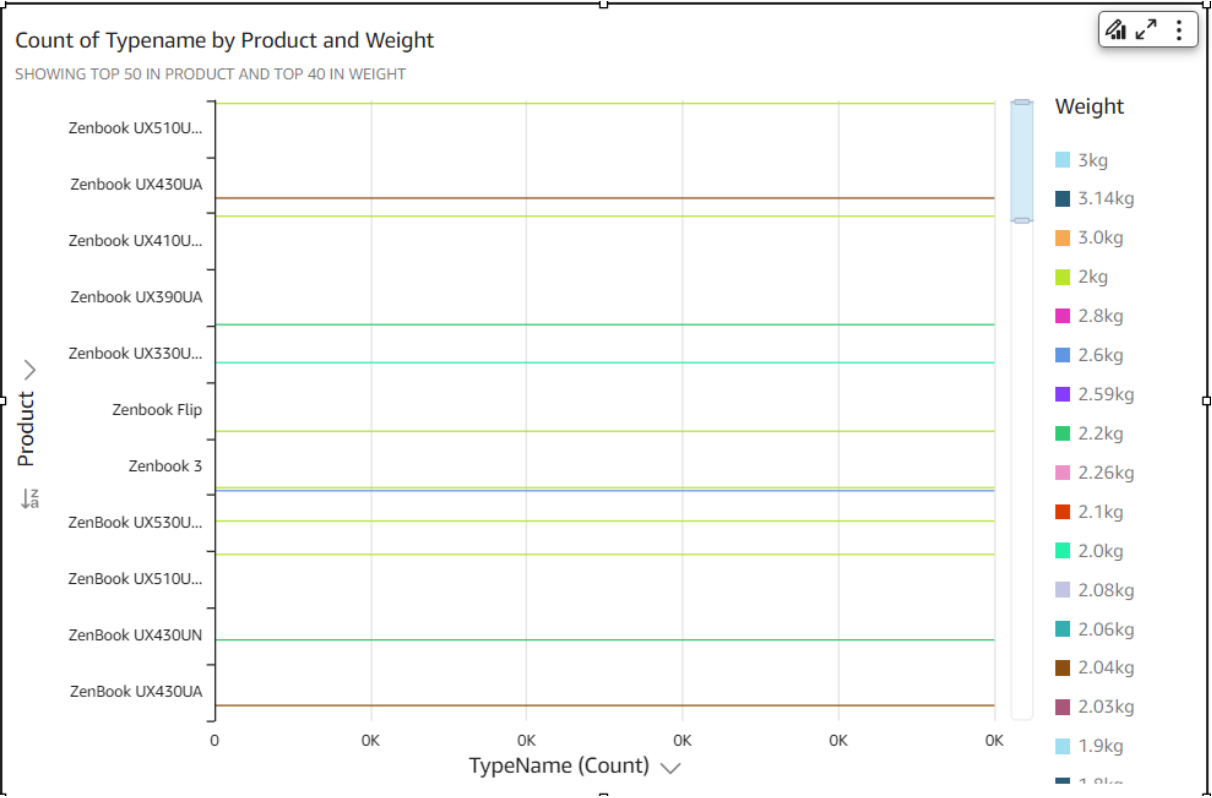


Hình 3. 9 Biểu đồ Phân Phối Giá của Các Nhà Sản Xuất



Hình 3. 10 Mối Quan Hệ giữa Kích Thước Màn Hình và Giá theo Tên loại

- Sử dụng biểu đồ cột để so sánh trọng lượng trung bình của các loại laptop.



Hình 3. 11 Biểu đồ so sánh trọng lượng trung bình của các loại laptop

## **CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU**

- Kết quả đạt được: Trong quá trình nghiên cứu về AWS QuickSight đã đạt được những kết quả tích cực về hiệu năng QuickSight cung cấp một giao diện trực quan và dễ sử dụng giúp người dùng dễ dàng truy cập một cách nhanh chóng.
- Công cụ này tích hợp mạnh mẽ với các dịch vụ khác nhau của AWS tạo điều kiện thuận lợi cho việc xử lý và phân tích dữ liệu từ nhiều nguồn khác nhau.

## **CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

- Tuy nhiên, cũng cần lưu ý rằng trực quan hóa dữ liệu chỉ là một công cụ, và nó không thể thay thế cho phân tích dữ liệu kỹ lưỡng. Chúng ta cần hiểu rõ về dữ liệu và các kỹ thuật trực quan hóa dữ liệu để có thể sử dụng trực quan hóa dữ liệu một cách hiệu quả.
- Trực quan hóa dữ liệu bằng công cụ AWS QuickSight là một nền tảng nghiên cứu đang phát triển nhanh chóng. Dưới đây là một số hướng phát triển tiềm năng của lĩnh vực này:
  - Phát triển các công cụ và kỹ thuật trực quan hóa dữ liệu mới.
  - Cải thiện hiệu quả của công việc trực quan hóa.
  - Tăng cường khả năng tương tác của việc trực quan hóa dữ liệu.

## **DANH MỤC TÀI LIỆU THAM KHẢO**

<https://aws.amazon.com/vi/what-is/data-visualization/>

<https://websitehcm.com/truc-quan-hoa-du-lieu/data-visualization/>

<https://aws.amazon.com/vi/quicksight/>

## **PHỤ LỤC**