**TRƯỜNG ĐẠI HỌC THUỶ LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**



Nhóm 18:

HỌ TÊN:

|  |  |
| --- | --- |
| Cao Thị Mai Phương | 2054015772 |
| Nguyễn Văn Luyện | 2151264669 |
| Hoàng Hiểu Nhi | 2154020987 |

**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN TRỰC QUAN HOÁ DỮ LIỆU**

**ĐỀ TÀI: TRỰC QUAN HOÁ NGƯỜI DÙNG NETFLIX**

**NGƯỜI HƯỚNG DẪN: Nguyễn Đắc Hiếu**

HÀ NỘI, năm 2024

Mục lục

[Phần 1: Giới thiệu bài toán 2](#_Toc3010)

[1.1. Giới thiệu tập dữ liệu 2](#_Toc20559)

[1.2. Lý thuyết các biểu đồ được dùng trong bài 3](#_Toc13685)

*[1.2.1. Biểu đồ đường (Line Chart)](#_Toc20303)* [3](#_Toc20303)

*[1.2.2. Biểu đồ Histogram](#_Toc16596)* [5](#_Toc16596)

*[1.2.3. Biểu đồ Boxplot](#_Toc7742)* [5](#_Toc7742)

*[1.2.4. Biểu đồ Countplot](#_Toc25217)* [6](#_Toc25217)

*[1.2.5. Biểu đồ mosaic plot](#_Toc26849)* [7](#_Toc26849)

*[1.2.6. Biểu đồ map](#_Toc25143)* [7](#_Toc25143)

*[1.2.7. Biểu đồ kdeplot](#_Toc32517)* [8](#_Toc32517)

*[1.2.8. Biểu đồ stackplot](#_Toc17309)* [9](#_Toc17309)

*[1.2.9. Biểu đồ histplot](#_Toc17529)* [10](#_Toc17529)

*[1.2.10. Biểu đồ heatmap](#_Toc14060)* [10](#_Toc14060)

[Phần 2: Trực quan hoá dữ liệu bằng biểu đồ 11](#_Toc7602)

[2.1. Phân tích tổng quan: Xu hướng người đăng ký netflix từ 04/2021 đến 01/2024 11](#_Toc6000)

*[- Nhận xét:](#_Toc9905)* [12](#_Toc9905)

[2.2. Số lượng người dùng phân phối theo độ tuổi 12](#_Toc12145)

[2.3. Phân bổ độ tuổi theo loại đăng ký 13](#_Toc11739)

*[- Nhận xét:](#_Toc29704)* [13](#_Toc29704)

[2.4. Số lượng người đăng ký và phân bổ doanh thu trên mỗi loại thiết bị 14](#_Toc31470)

*[- Lý do chọn biểu đồ:](#_Toc16614)* [14](#_Toc16614)

*[- Nhận xét:](#_Toc27187)* [14](#_Toc27187)

[2.5. So sánh các thiết bị sử dụng trên từng loại đăng kí 15](#_Toc6943)

*[- Lý do chọn biểu đồ:](#_Toc26261)* [15](#_Toc26261)

*[- Nhận xét:](#_Toc31600)* [15](#_Toc31600)

[2.6. Tổng doanh thu theo quốc gia 16](#_Toc24922)

*[- Lý do chọn biểu đồ:](#_Toc25501)* [16](#_Toc25501)

[2.7. So sánh số lượng người dùng sử dụng các loại đăng ký theo độ tuổi 16](#_Toc7292)

*[- Lý do chọn biểu đồ:](#_Toc24539)* [17](#_Toc24539)

*[- Nhận xét:](#_Toc1274)* [17](#_Toc1274)

[2.8. Tương quan giữa các thuộc tính 18](#_Toc23646)

*[- Lý do chọn biểu đồ:](#_Toc1183)* [18](#_Toc1183)

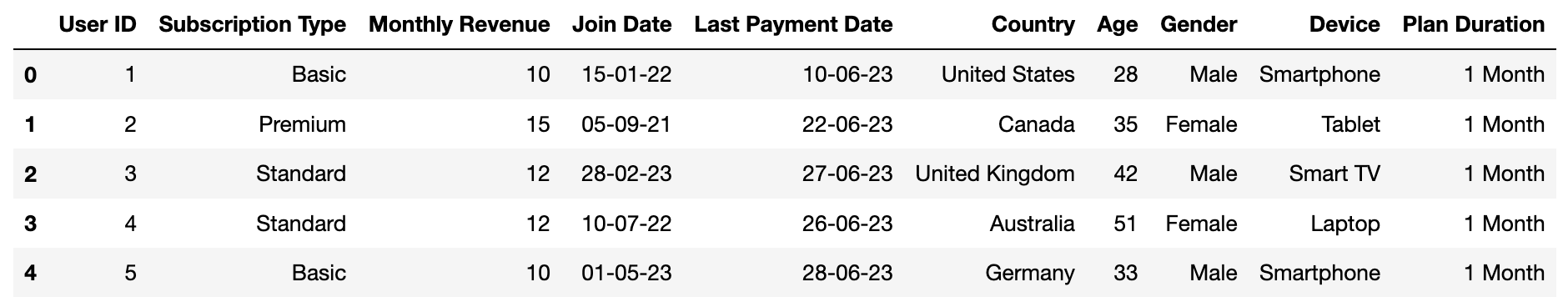
*[- Nhận xét:](#_Toc14467)* [18](#_Toc14467)

# Phần 1: Giới thiệu bài toán

## 1.1. Giới thiệu tập dữ liệu

Tên bài toán: Trực quan hoá dữ liệu người dùng Netflix.

Tập dữ liệu gồm: 2500 dòng và 10 cột.



Tập dữ liệu này cung cấp về cơ sở người dùng Netflix mẫu, hiển thị các khía cạnh khác nhau về đăng ký, doanh thu, chi tiết tài khoản và hoạt động của người dùng. Mỗi hàng đại diện cho một người dùng duy nhất, được xác định bằng ID người dùng của họ. Tập dữ liệu bao gồm các thông tin như loại đăng ký của người dùng (Cơ bản, Tiêu chuẩn hoặc Cao cấp), doanh thu hàng tháng được tạo từ đăng ký của họ, ngày họ tham gia Netflix (Ngày tham gia), ngày thanh toán cuối cùng của họ (Ngày thanh toán cuối cùng) và quốc gia mà họ đang ở.

Các cột bổ sung đã được đưa vào để cung cấp thông tin chi tiết về hành vi và sở thích của người dùng. Các cột này bao gồm Loại thiết bị (ví dụ: Smart TV, Điện thoại di động, Máy tính để bàn, Máy tính bảng) và Trạng thái tài khoản (cho dù tài khoản có đang hoạt động hay không). Tập dữ liệu đóng vai trò là bản trình bày tổng hợp và không phản ánh dữ liệu người dùng Netflix thực tế. Nó có thể được sử dụng để phân tích và lập mô hình nhằm hiểu xu hướng, sở thích của người dùng và tạo doanh thu trong cơ sở người dùng Netflix giả định.

Dữ Liệu gồm các cột như:

|  |  |
| --- | --- |
| **Tên thuộc tính** | **Ý nghĩa** |
| User ID | Là một định danh duy nhất cho mỗi người dùng. |
| Subscription Type | Loại đăng ký của người dùng, có thể là miễn phí, trả phí hoặc có các tính năng khác nhau. |
| Monthly Revenue | Số tiền thu được từ mỗi người dùng trong mỗi tháng. |
| Join Date | Ngày mà người dùng đã tham gia vào dịch vụ hoặc sản phẩm. |
| Last Payment Date | Ngày mà người dùng đã thực hiện thanh toán gần nhất. |
| Country | Quốc gia người dùng đăng ký. |
| Age | Tuổi của người dùng. |
| Gender | Giới tính |
| Device | Thiết bị mà người dùng sử dụng để truy cập vào dịch vụ hoặc sản phẩm. |
| Plan Duration | Thời gian mà người dùng đã đăng ký hoặc đặt mua dịch vụ hoặc sản phẩm. |

## 1.2. Lý thuyết các biểu đồ được dùng trong bài

### *1.2.1. Biểu đồ đường (Line Chart)*

Biểu đồ đường (Line Chart) là biểu đồ thể hiện mức dao động giá và khối lượng giao dịch dưới dạng một đường liền nét.

#### *1.2.1.1. Đặc điểm:*

* Biểu đồ thường được sử dụng cho những thị trường chứng khoán. Cung cấp cho các nhà giao dịch một hình dung rõ ràng về nơi mà giá của chứng khoán đã đi trong khoảng thời gian nhất định.
* Biểu đồ thể hiện xu hướng: Biểu đồ đường giúp người dùng xem được xu hướng tăng hoặc giảm của dữ liệu theo thời gian. bẰng cách quan sát các đường di chuyển trên biểu đồ mà người dùng có thể nhận ra các mô hình và xu hướng chung.
* So sánh dữ liệu: Có thể so sánh dữ liệu khác nhau trên cùng biểu đồ. Điều này giúp nhận ra mối quan hệ và sự khác biệt giữa các dữ liệu.
* Dự báo và dự đoán: Biểu đồ đường có có thể được sử dụng để dự đoán và dự báo sự phát triển trong tương lai của dữ liệu dựa trên xu hướng hiện tại.

#### *1.2.1.2. Ưu điểm và hạn chế*

Biểu đồ đường (Line chart) trong việc trình bày dữ liệu số liệu thống kê có nhiều ưu điểm:

* Thể hiện xu hướng và biến động: Line chart giúp người dùng dễ dàng nhìn thấy xu hướng và biến động của dữ liệu qua thời gian hoặc giá trị liên tục. Việc kết nối các điểm dữ liệu bằng đường thẳng liên tục giúp người dùng nhận ra các mô hình, sự tăng giảm và sự tương quan giữa các giá trị.
* Trực quan và dễ hiểu: Việc trình bày và đọc dữ liệu trở lên trực quan và dễ hiểu hơn. Đường thẳng trên biểu đồ giúp người xem dễ dàng theo dõi sự thay đổi của dữ liệu và phân tích mô hình một cách nhanh chóng.
* Phù hợp cho dữ liệu số liệu thống kê: Line chart thích hợp cho việc biểu diễn dữ liệu số liệu thống kê như doanh thu theo thời gian, số lượng truy cập trang web theo ngày, chỉ số tài chính và nhiều biến số khác. Nó giúp người dùng dễ dàng nhìn thấy sự biến đổi và so sánh giữa các giá trị.
* Đa dạng và linh hoạt: Line chart có thể được tùy chỉnh và định dạng theo nhu cầu của người dùng. Bạn có thể tùy chỉnh màu sắc, kiểu đường, nhãn trục và các thuộc tính khác để tạo ra biểu đồ theo ý muốn.
* Dễ so sánh: Line chart cho phép người dùng so sánh nhanh chóng giữa các chuỗi dữ liệu hoặc phân tích sự khác biệt giữa các dòng dữ liệu. Điều này giúp người dùng rút ra những nhận định và quyết định dựa trên sự so sánh dễ dàng.
* Tích hợp với các công cụ phân tích: Line chart thường được sử dụng trong các công cụ phân tích dữ liệu như báo cáo tài chính, công cụ BI (Business Intelligence) và các phần mềm thống kê. Việc sử dụng line chart giúp người dùng phân tích, dự đoán và đưa ra quyết định dựa trên dữ liệu số thống kê.

Tuy nhiên, cần lưu ý rằng line chart không phù hợp cho các dạng dữ liệu phi số hoặc không liên tục. Đối với những tình huống như vậy, có thể sử dụng các loại biểu đồ khác như bar chart hoặc pie chart.

### ***1.2.2. Biểu đồ Histogram***

#### *1.2.2.1. Đặc điểm*

Biểu đồ Histogram là một dạng biểu đồ thể hiện tần suất theo dạng cột. Dữ liệu được biểu thị bằng các cột trên biểu đồ có độ cao khác nhau tùy thuộc vào tần suất (bao nhiêu lần) phạm vi dữ liệu cụ thể xảy ra.

#### *1.2.2.2. Ưu điểm và nhược điểm*

Ưu điểm chính của biểu đồ này là tính đơn giản và tính linh hoạt của nó. Nó có thể được sử dụng trong nhiều tình huống khác nhau để cung cấp một cái nhìn sâu sắc về phân bố số liệu. Ví dụ, nó có thể được sử dụng trong bán hàng và tiếp thị để phát triển các kế hoạch định giá và chiến dịch tiếp thị hiệu quả nhất.

Khi kích thước mẫu nhỏ, có quá ít điểm dữ liệu thì Histogram không thể hiển thị sự phân phối của dữ liệu. Theo quy tắc, biểu đồ tần suất hữu ích khi có 20 hoặc nhiều quan sát hơn.

Khi có khoảng trống lớn trong dữ liệu. Biểu đồ Histogram thích hợp nhất khi dữ liệu được thu thập liên tục. Vì vậy khi thiếu hoặc không xác định biểu đồ không thể phát huy được hiệu quả của mình.

### ***1.2.3. Biểu đồ Boxplot***

#### *1.2.3.1: Đặc điểm*

* Biểu đồ Boxplot là biểu đồ diễn tả 5 vị trí phân bố của dữ liệu, đó là: giá trị nhỏ nhất (min), tứ phân vị thứ nhất (Q1), trung vị (median), tứ phân vị thứ 3 (Q3) và giá trị lớn nhất (max).

#### *1.2.3.2: Ưu điểm và nhược điểm*

Ưu điểm:

* Hiển thị phân phối: Box plot cho phép bạn nhanh chóng nhận biết phân phối của dữ liệu, bao gồm giá trị trung bình, phương sai, độ lệch và sự phân tán.
* So sánh giữa các nhóm: Box plot cho phép so sánh phân phối của các nhóm dữ liệu khác nhau trên cùng một biểu đồ. Điều này giúp dễ dàng so sánh trực quan các đặc điểm thống kê của các nhóm.
* Phát hiện ngoại lệ: Box plot giúp phát hiện nhanh chóng sự xuất hiện của các giá trị ngoại lệ trong dữ liệu. Các giá trị ngoại lệ thường được biểu thị như các điểm nằm ngoài phạm vi của các "whisker" của box plot.
* Khái quát và đơn giản: Box plot là một công cụ trực quan đơn giản và dễ hiểu, phù hợp để trình bày dữ liệu một cách rõ ràng và tóm tắt.

Nhược điểm:

* Mất thông tin chi tiết: Mặc dù box plot cung cấp một cái nhìn tổng quan về phân phối của dữ liệu, nhưng nó không cung cấp thông tin chi tiết về hình dạng chính xác của phân phối như histogram.
* Không phản ánh kích thước mẫu: Box plot không thể hiện kích thước của mỗi nhóm dữ liệu, do đó, không cung cấp thông tin về độ tin cậy của ước lượng thống kê.
* Không thể thể hiện tương quan: Box plot không thể hiện mối quan hệ tương quan giữa các biến, mà chỉ thể hiện phân phối của mỗi biến một cách độc lập.

### ***1.2.4. Biểu đồ Countplot***

#### *1.2.4.1. Đặc điểm*

* Countplot là một biểu đồ thường được sử dụng trong khoa học dữ liệu để đếm số lượng quan sát trong mỗi nhóm và trực quan hóa chúng dưới dạng cột dọc. Dưới đây là một số ưu điểm và nhược điểm của countplot:

#### *1.2.4.2. Ưu điểm và nhược điểm*

Ưu điểm:

* Dễ hiểu: Countplot là một biểu đồ đơn giản và dễ hiểu. Nó chỉ đơn giản là đếm và trực quan hóa số lượng quan sát trong từng nhóm.
* Phân loại dữ liệu rời rạc: Countplot là lựa chọn tốt khi bạn muốn phân loại dữ liệu rời rạc, như các biến phân loại (categorical variables).
* Tích hợp tốt với dữ liệu lớn: Khi bạn có một lượng lớn dữ liệu, countplot vẫn có thể trực quan hóa thông tin một cách dễ dàng.
* Sử dụng cho so sánh: Countplot thích hợp để so sánh số lượng quan sát giữa các nhóm khác nhau, giúp nhận diện sự khác biệt một cách trực quan.

Nhược điểm:

* Hạn chế với dữ liệu liên tục: Countplot không phù hợp cho dữ liệu liên tục, vì nó chỉ thực hiện đếm và trực quan hóa số lượng quan sát trong từng nhóm.
* Mất đi thông tin chi tiết: Trong một số trường hợp, countplot có thể mất đi một số thông tin chi tiết, đặc biệt là khi có quá nhiều nhóm hoặc khi số lượng quan sát trong mỗi nhóm quá nhỏ.
* Khó khăn khi so sánh tỷ lệ: Countplot không thích hợp để so sánh tỷ lệ giữa các nhóm, mà chỉ phù hợp cho so sánh số lượng quan sát.
* Cần chú ý đến biểu đồ lệch: Nếu dữ liệu của bạn có sự lệch nghiêng (skewed), countplot có thể không phản ánh chính xác sự phân bố của dữ liệu.

### ***1.2.5. Biểu đồ mosaic plot***

#### *1.2.5.1. Đặc điểm*

* Biểu đồ Mosaic, còn được gọi là biểu đồ mặt cắt (mosaic plot), là một công cụ trực quan hóa dữ liệu phân loại (categorical data). Nó biểu diễn mối quan hệ giữa các biến phân loại bằng cách phân chia kích thước của các hình chữ nhật hoặc hình vuông theo tỷ lệ với tổng lượng quan sát. Đây là một cách hiệu quả để trực quan hóa dữ liệu đa biến.

#### *1.2.5.2. Ưu điểm và nhược điểm*

Ưu điểm của biểu đồ Mosaic:

* Trực quan và Dễ hiểu: Biểu đồ Mosaic cung cấp một cách trực quan và dễ hiểu để thể hiện mối quan hệ giữa các biến phân loại.
* Hiển thị Tích chập của Dữ liệu: Nó cho phép bạn nhìn thấy cả mối quan hệ giữa biến chính và biến phụ, cũng như phân bố của cả hai.
* Tích hợp được với các biến phân loại nhiều cấp độ: Biểu đồ Mosaic thích hợp cho việc trực quan hóa mối quan hệ giữa các biến phân loại có nhiều cấp độ.
* Khả năng so sánh: Nó cho phép so sánh mối quan hệ giữa các nhóm của các biến phân loại một cách dễ dàng.

Nhược điểm của biểu đồ Mosaic:

* Khó đọc với số lượng nhóm lớn: Nếu có quá nhiều nhóm, biểu đồ Mosaic có thể trở nên rối mắt và khó đọc.
* Hạn chế với dữ liệu lớn: Đối với các tập dữ liệu lớn, biểu đồ Mosaic có thể trở nên quá phức tạp và khó hiểu.
* Không thể phân tích tương quan: Mặc dù biểu đồ Mosaic thể hiện mối quan hệ giữa các biến phân loại, nhưng nó không cung cấp thông tin về mức độ tương quan giữa chúng.ung cấp thông tin về mức độ tương quan giữa chúng.

### ***1.2.6. Biểu đồ map***

#### *1.2.6.1. Đặc điểm*

* Biểu đồ map, hay bản đồ trực quan (map visualization), là một công cụ mạnh mẽ để trực quan hóa dữ liệu địa lý và hiển thị thông tin trên một bản đồ địa lý. Dưới đây là một số ưu điểm và nhược điểm của việc sử dụng biểu đồ map:

#### *1.2.6.2. Ưu điểm và nhược điểm*

Ưu điểm của biểu đồ map:

* Trực quan hóa dữ liệu địa lý: Biểu đồ map cho phép bạn hiển thị thông tin dữ liệu trên một bản đồ địa lý, giúp bạn hiểu rõ hơn về mối quan hệ không gian của dữ liệu.
* Phát hiện mẫu không gian: Bằng cách trực quan hóa dữ liệu địa lý, bạn có thể phát hiện ra các mẫu không gian và xu hướng phân bố của dữ liệu, giúp bạn đưa ra những quyết định phù hợp.
* Hiểu biết sâu sắc về vị trí: Biểu đồ map giúp bạn hiểu rõ hơn về vị trí địa lý của các sự kiện, đối tượng hoặc hiện tượng mà bạn quan tâm.
* Tích hợp với dữ liệu đa dạng: Bạn có thể tích hợp nhiều loại dữ liệu khác nhau trên cùng một bản đồ, từ dữ liệu địa lý cơ bản đến dữ liệu phức tạp như dữ liệu thống kê hay dữ liệu thời gian.
* Giao tiếp hiệu quả: Biểu đồ map là một cách hiệu quả để giao tiếp thông tin phức tạp và phân tích dữ liệu địa lý cho mọi người, kể cả những người không có kiến thức chuyên sâu về thống kê hay khoa học dữ liệu.

Nhược điểm của biểu đồ map:

* Biased và biểu đồ lỗi: Nếu không sử dụng dữ liệu địa lý chính xác hoặc không chú ý đến bias, biểu đồ map có thể dẫn đến sự hiểu lầm hoặc phản ứng không chính xác từ phía người xem.
* Phần mềm và dữ liệu phức tạp: Đôi khi, việc tạo và hiệu chỉnh biểu đồ map có thể phức tạp và đòi hỏi sự hiểu biết về phần mềm GIS (hệ thống thông tin địa lý) và các kỹ thuật trực quan hóa dữ liệu địa lý.
* Giới hạn của việc trực quan hóa địa lý: Trong một số trường hợp, việc trực quan hóa dữ liệu địa lý không thể hiển thị được mối quan hệ phức tạp giữa các biến hoặc không thể hiện được thông tin chi tiết đến từng điểm dữ liệu.
* Phụ thuộc vào phân tích thêm: Trong nhiều trường hợp, biểu đồ map cần được kết hợp với phân tích thống kê hoặc dữ liệu bổ sung để hiểu rõ hơn về mối quan hệ giữa các biến.

### ***1.2.7. Biểu đồ kdeplot***

#### *1.2.7.1. Đặc điểm*

* KDEplot (Kernel Density Estimate plot) là một loại biểu đồ thống kê dùng để ước lượng hàm mật độ xác suất (probability density function) của một biến ngẫu nhiên. Nó thường được sử dụng để hiển thị phân bố xác suất của dữ liệu và tạo ra một ước lượng mượt mà của hàm mật độ xác suất từ các điểm dữ liệu.

#### *1.2.7.2. Ưu đểm và nhược điểm*

Ưu điểm của KDEplot:

* Trực quan hóa phân phối dữ liệu: KDEplot cung cấp một cách trực quan để hiểu phân phối của dữ liệu. Bằng cách trực quan hóa hàm mật độ xác suất, nó cho phép bạn nhìn thấy cách dữ liệu phân phối một cách mượt mà và linh hoạt hơn so với việc sử dụng các biểu đồ khác như histogram.
* Dễ so sánh giữa các nhóm: Nếu bạn có nhiều nhóm dữ liệu, bạn có thể dễ dàng so sánh hàm mật độ xác suất của chúng bằng cách vẽ chúng trên cùng một biểu đồ.
* Phát hiện mẫu và cấu trúc: KDEplot có thể giúp bạn phát hiện mẫu và cấu trúc trong dữ liệu một cách trực quan. Bằng cách xem xét độ dày của hàm mật độ xác suất tại các vị trí khác nhau, bạn có thể nhận biết mối quan hệ và cấu trúc phức tạp trong dữ liệu.

Nhược điểm của KDEplot:

* Đòi hỏi lượng dữ liệu lớn: Để có một ước lượng chính xác của hàm mật độ xác suất, bạn cần có một lượng dữ liệu đủ lớn. Điều này có nghĩa là KDEplot không phù hợp với các tập dữ liệu nhỏ.
* Độ phức tạp tính toán: Trong một số trường hợp, việc tính toán hàm mật độ xác suất có thể tốn nhiều thời gian và tài nguyên tính toán.
* Khó hiểu với người không chuyên: Đối với những người không quen thuộc với khái niệm của hàm mật độ xác suất, KDEplot có thể khó hiểu và cần sự giải thích.

### ***1.2.8. Biểu đồ stackplot***

Stackplot là biểu đồ thường được sử dụng để hiển thị sự biến đổi của các thành phần riêng lẻ trong một tổng thể qua thời gian hoặc theo các biến số khác. Dưới đây là một số đặc điểm và ưu nhược điểm của stackplot.

#### *1.2.8.1. Đặc điểm*

Hiển thị phân phối tổng thể: Stackplot cho phép người xem nhận thấy tổng thể của một tập hợp các biến số được phân chia thành các thành phần riêng lẻ.

Biểu diễn sự biến đổi qua thời gian hoặc biến số: Stackplot thường được sử dụng để biểu diễn sự biến đổi của các thành phần trong một tổng thể qua thời gian hoặc theo các biến số khác nhau.

Dễ so sánh: Stackplot cho phép người xem dễ dàng so sánh sự biến đổi của các thành phần trong tổng thể.

#### *1.2.8.2. Ưu điểm và nhược điểm*

Ưu điểm:

Trực quan: Stackplot tạo ra một biểu đồ trực quan và dễ hiểu, giúp người xem nhanh chóng nhận biết sự biến đổi của các thành phần.

Hiển thị tỷ lệ phần trăm: Stackplot thường cho phép hiển thị tỷ lệ phần trăm của mỗi thành phần so với tổng thể, giúp làm rõ hơn phần trăm đóng góp của từng thành phần.

Phản ánh sự biến đổi tổng thể: Stackplot giúp người xem nhận biết sự biến đổi của tổng thể dựa trên sự biến đổi của các thành phần.

Nhược điểm:

Khó phân tích với số lượng thành phần lớn: Khi có quá nhiều thành phần, stackplot có thể trở nên rối mắt và khó phân tích.

Yêu cầu dữ liệu phải được chia thành các thành phần rõ ràng: Để sử dụng stackplot một cách hiệu quả, dữ liệu cần phải được chia thành các thành phần rõ ràng và có tính nhất quán.

Khó khăn trong việc so sánh giữa các biến số không liên tục: Stackplot không phù hợp để so sánh giữa các biến số không liên tục hoặc không đồng nhất về mặt thời gian hoặc biến số khác.

### ***1.2.9. Biểu đồ histplot***

#### *1.2.9.1.Đặc điểm*

* Biểu đồ Histplot (viết tắt là Histogram plot) là dạng biểu đồ thống kê được phát triển dựa trên biểu đồ Histogram truyển thống.Kết hợp Histogram và KDE (Kernel Density Estimation): Hiển thị đồng thời cả phân bố tần suất (Histogram) và mật độ phân bố (KDE) của dữ liệu trên cùng một biểu đồ.

#### *1.2.9.2. Ưu và nhược điểm*

Ưu điểm:

* Cung cấp thông tin chi tiết hơn: kết hợp cả biểu đồ histogram và KDE giúp ta có cái nhìn tổng quát hơn về phân bố dữ liệu, bao gồm cả hình dạng và mật độ phân bố như: sự đối xứng, sự tập trung, sự dẹt, sự hiện diện các giá trị bất thường.
* Dễ so sánh: Cho phép so sánh trực quan phân bố của nhiều tập dữ liệu khác nhau trên cùng một biểu đồ.
* Linh hoạt: Khả năng tùy chỉnh cao giúp ta tạo ra những biểu đồ phù hợp với nhu cầu và mục đích cụ thể.

Nhược điểm:

* Tương đối mới: Biểu đồ Histplot là một phát minh tương đối mới so với Histogram truyền thống. Do đó, nó chưa được sử dụng phổ biến và có thể ít được biết đến hơn.
* Yêu cầu kiến thức: Việc sử dụng biểu đồ Histplot hiệu quả đòi hỏi người dùng có kiến thức nhất định về thống kê và phân tích dữ liệu. Nhờ vậy, ta có thể hiểu rõ ý nghĩa của các thông tin được hiển thị trên biểu đồ và sử dụng chúng để đưa ra các phân tích và đánh giá chính xác.
* Có thể khó hiểu: Đối với người mới bắt đầu, việc kết hợp Histogram và KDE trong biểu đồ Histplot có thể khiến họ khó hiểu và khó khăn trong việc giải thích các thông tin được hiển thị.
* Tốn nhiều thời gian: Việc tạo biểu đồ Histplot với nhiều tùy chỉnh có thể tốn nhiều thời gian hơn so với Histogram truyền thống.

### ***1.2.10. Biểu đồ heatmap***

Biểu đồ heatmap là một công cụ mạnh mẽ để hiển thị dữ liệu theo một ma trận màu sắc, trong đó màu sắc của từng ô đại diện cho giá trị của một điểm dữ liệu. Dưới đây là các đặc điểm, ưu điểm và nhược điểm của biểu đồ heatmap:

#### *1.2.10.1. Đặc điểm*

Mã hóa màu sắc: Dữ liệu được mã hóa bằng màu sắc, giúp người xem dễ dàng nhận biết các giá trị lớn, nhỏ hoặc trung bình một cách trực quan.

Hiển thị mối quan hệ: Heatmap giúp hiển thị mối quan hệ giữa các biến số trong một ma trận dữ liệu. Điều này có thể giúp bạn phát hiện các mẫu hoặc sự liên quan giữa các biến số.

Dễ so sánh: Bằng cách sử dụng màu sắc, người xem có thể dễ dàng so sánh giữa các giá trị trong cùng một hàng hoặc cột.

Phù hợp cho dữ liệu lớn: Heatmap có thể xử lý được các ma trận dữ liệu lớn một cách hiệu quả, làm cho nó phù hợp cho việc hiển thị dữ liệu phức tạp.

#### *1.2.10.2. Ưu và nhược điểm*

Ưu điểm:

Trực quan hóa dữ liệu phức tạp: Heatmap giúp trực quan hóa dữ liệu phức tạp một cách dễ hiểu và hấp dẫn.

Dễ tìm kiếm thông tin: Với heatmap, người xem có thể dễ dàng tìm kiếm thông tin cụ thể bằng cách quét qua ma trận màu sắc.

Dễ tùy chỉnh: Heatmap có thể được tùy chỉnh để thể hiện các giá trị khác nhau và màu sắc theo ý muốn.

Nhược điểm:

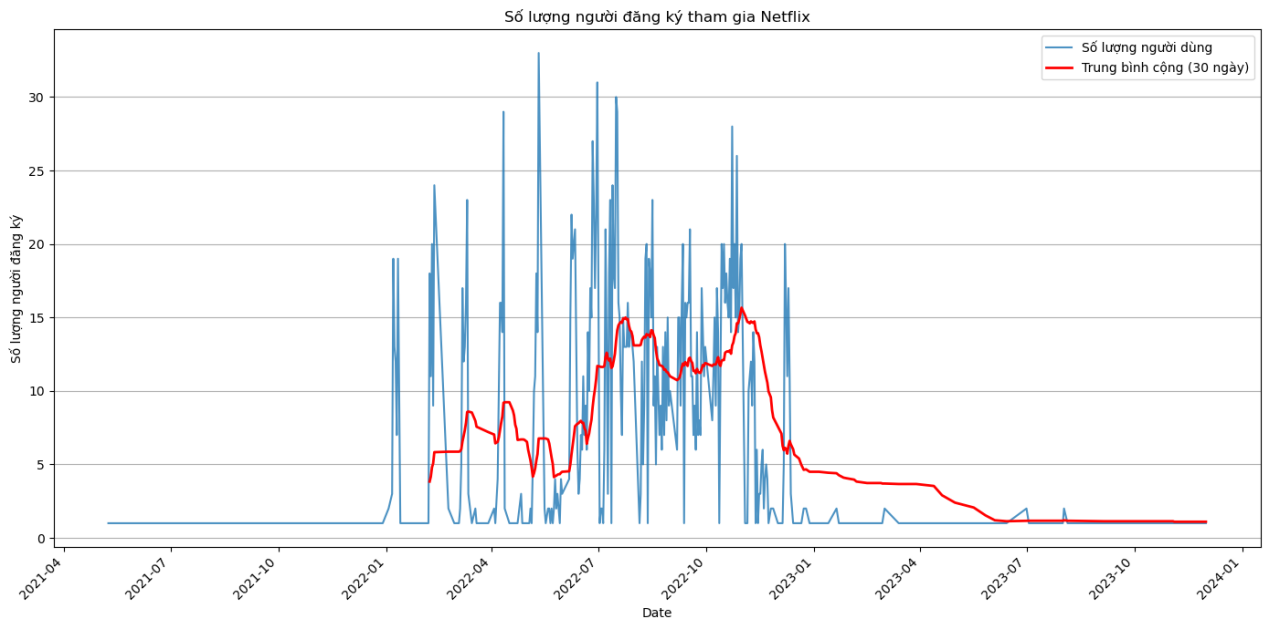
Mất thông tin chi tiết: Heatmap có thể mất đi các chi tiết khi dữ liệu quá lớn hoặc khi số lượng ô trên biểu đồ quá nhiều.

Dễ bị hiểu nhầm: Mặc dù heatmap là một công cụ trực quan, nhưng cũng có thể dễ bị hiểu nhầm nếu không có thông tin bổ sung về ý nghĩa của mỗi giá trị màu sắc.

Không phản ánh sự biến thiên: Heatmap thường không phản ánh được sự biến thiên của dữ liệu trong không gian chiều, mà chỉ thể hiện giá trị tại mỗi điểm dữ liệu.

# Phần 2: Trực quan hoá dữ liệu bằng biểu đồ

## 2.1. Phân tích tổng quan: Xu hướng người đăng ký netflix từ 04/2021 đến 01/2024



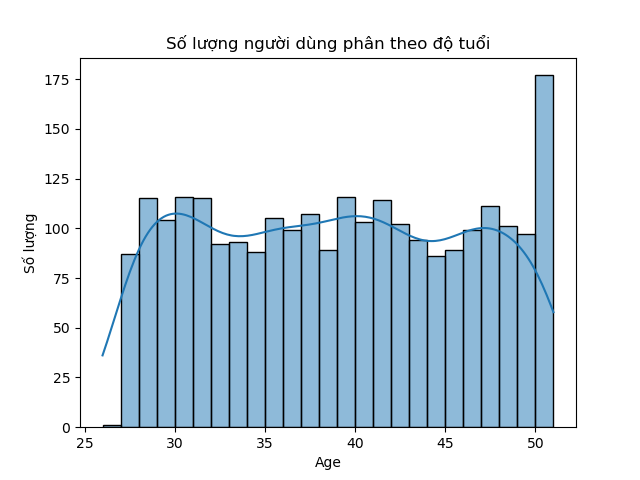
Mô tả: Biểu đồ thể hiện xu hướng người dùng đăng ký netflix từ tháng 04/2021 đến 01/2024

***- Lý do chọn biểu đồ:*** Biểu đồ đường thể hiện được sự biến dữ liệu của thời gian liên tục. Giúp người đọc nhìn thấy rõ được xu hướng và dễ dàng nhận xét sự tăng giảm của số lượng người dùng.

***- Nhận xét:***

Dữ liệu bắt đầu từ tháng 4 năm 2021 và kết thúc vào cuối năm 2023. Số lượng người dùng đăng ký có sự biến động mạnh từ tháng 1 năm 2022 đến tháng 1 năm 2022. Đây là khoảng thời gian số lượng người đăng ký sử dụng Netflix tăng cao. Do ảnh hưởng từ đại dịch COVID-19 mọi người thường xuyên ở nhà vì vậy số lượng người đăng ký ở trong giai đoạn này tăng cao so với các giai đoạn khác, đỉnh điểm là từ tháng 8 năm 2022 đến tháng 11 cùng năm. Từ tháng 11 năm 2022 đến tháng 1 năm 2023 số lượng người đăng ký giảm mạnh nhưng vẫn chưa về thời điểm bình ổn ban đầu và giữ ở mức trung bình đến tháng 4 năm 2023 số lượng người đăng ký lại tiếp tục giảm đến đầu tháng 7 năm 2023 trở đi thì số lượng người đăng ký Netflix đã trở lại và giữ ở mức ban đầu.

## 2.2. Số lượng người dùng phân phối theo độ tuổi



Mô tả:

Biểu đồ Histogram thể hiện số lượng người dùng theo độ tuổi

***- Lý do chọn biểu đồ:***

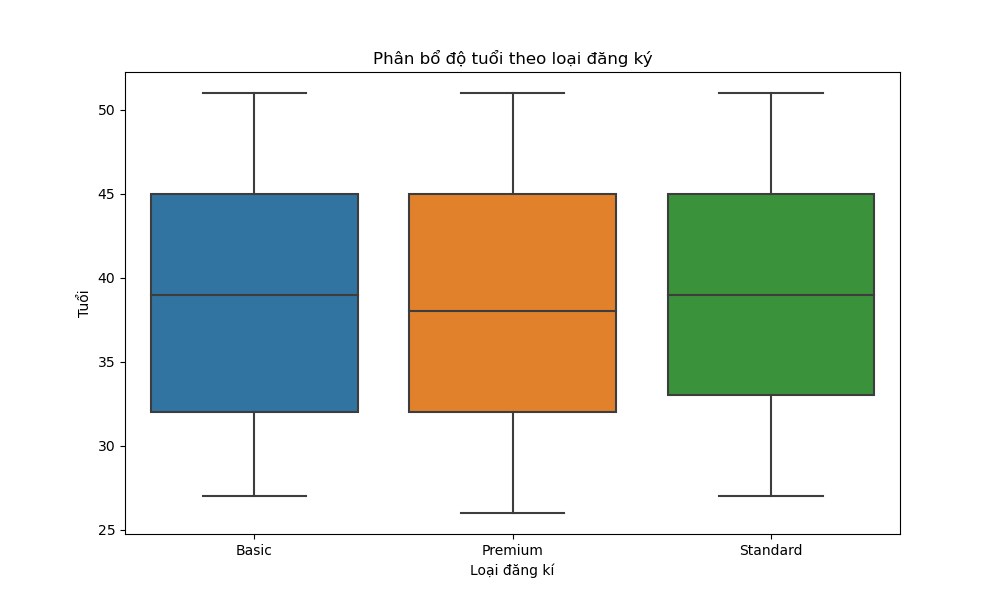
Hiển thị trực quan: Biểu đồ Histogram thể hiện sự phân bố tần suất của số lượng người dùng dưới dạng các cột, giúp người xem dễ dàng hình dung sự tập trung và biến động của giá. So với các dạng biểu đồ khác như biểu đồ hình hộp hay biểu đồ mật độ, biểu đồ Histogram đơn giản và dễ hiểu hơn, phù hợp với nhiều đối tượng người dùng.

Biểu đồ Histogram cho phép chia nhỏ dữ liệu số lượng người dùng thành các nhóm (bin) cụ thể, giúp người xem phân tích chi tiết mức độ tập trung của nó trong từng nhóm.

Thông qua độ cao của các cột, người xem có thể so sánh tần suất xuất hiện của các độ tuổi khác nhau, từ đó xác định các giá trị phổ biến và ít phổ biến.

***- Nhận xét:*** Độ tuổi đăng ký sử dụng Netflix nằm trong khoảng từ 27 tuổi đến 51 tuổi. Số lượng người đăng ký không có sự chênh lệch nhiều từ độ tuổi 28-50. Tuy nhiên số lượng người đăng ký ở độ tuổi 50 tuổi là rất cao, cao gần như gấp đôi so với các độ tuổi còn lại. Độ tuổi có số người đăng ký thấp nhất là 27 tuổi.

## 2.3. Phân bổ độ tuổi theo loại đăng ký



Mô tả:Biểu đồ boxplot thể hiện sự phân bố theo độ tuổi dựa vào loại đăng ký dịch vụ

***- Lý do chọn biểu đồ:*** Biểu đồ boxplot cho phép so sánh sự phân phối giữa các nhóm dữ liệu với nhau trên cùng một biểu đồ. Hộp thể hiện giá trị trung vị, tứ phân vị thứ nhất và tứ phân vị thứ 3; Râu thể hiện giá trị tối đa và giá trị tối thiểu.

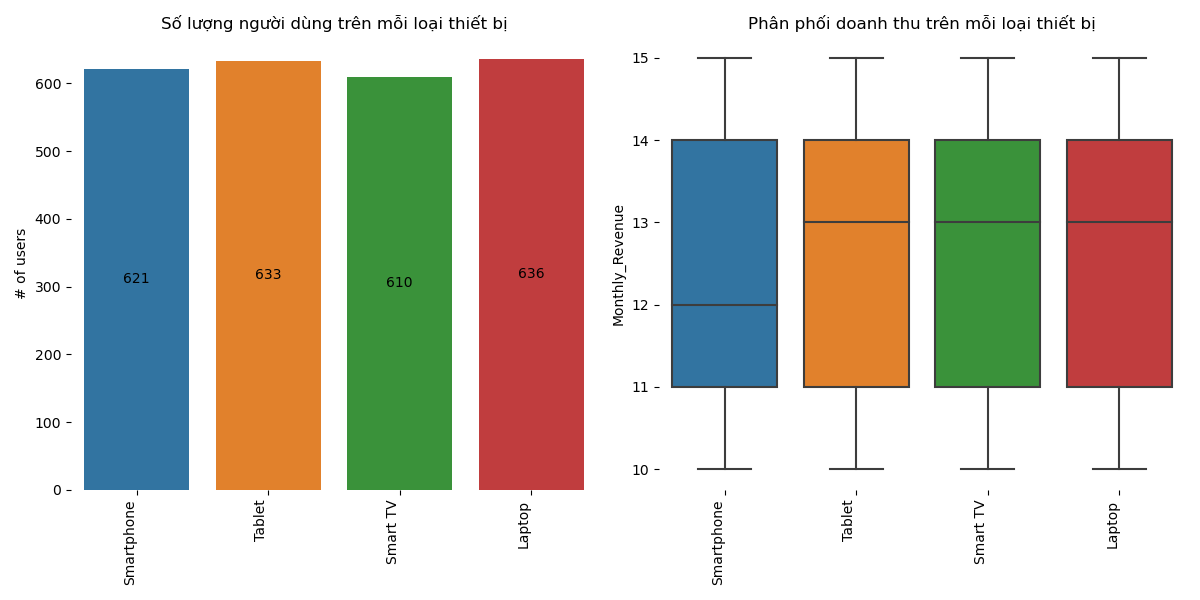
***- Nhận xét:***

Sự phân bố người dùng : không đối xứng

Phạm vi người dùng ở mỗi loại đăng ký gần như là như nhau tương đối rộng, từ độ tuổi 27 đến 52 tuổi. Chỉ có Loại đăng ký Premium là thấp nhất là 26 tuổi.

Cả ba loại đăng ký tập trung vào độ tuổi gần như là như nhau, chỉ có loại đăng ký Standard tuổi của người dùng sử dụng cao hơn hai loại còn lại 1 tuổi không đáng kể. Ở loại đăng ký Basic thì median lệch phía Q3 nhiều hơn => Loại đăng ký Basic có nhiều người dùng lớn tuổi hơn người dùng ít tuổi. Hai loại đăng ký còn lại thì mức độ phân bổ trên các độ tuổi ở trạng thái bình thường.

## 2.4. Số lượng người đăng ký và phân bổ doanh thu trên mỗi loại thiết bị



Biểu đồ countplot: thể hiện số lượng người dùng trên mỗi loại thiết bị

Biểu đồ boxplot: thể hiện phân phối doanh thu trên mỗi loại thiết bị

***- Lý do chọn biểu đồ:***

Biểu đồ countplot: Biểu đồ biểu thị được số lượng mỗi nhóm theo cột và sử dụng cho kiểu dữ liệu phân loại (categorical).

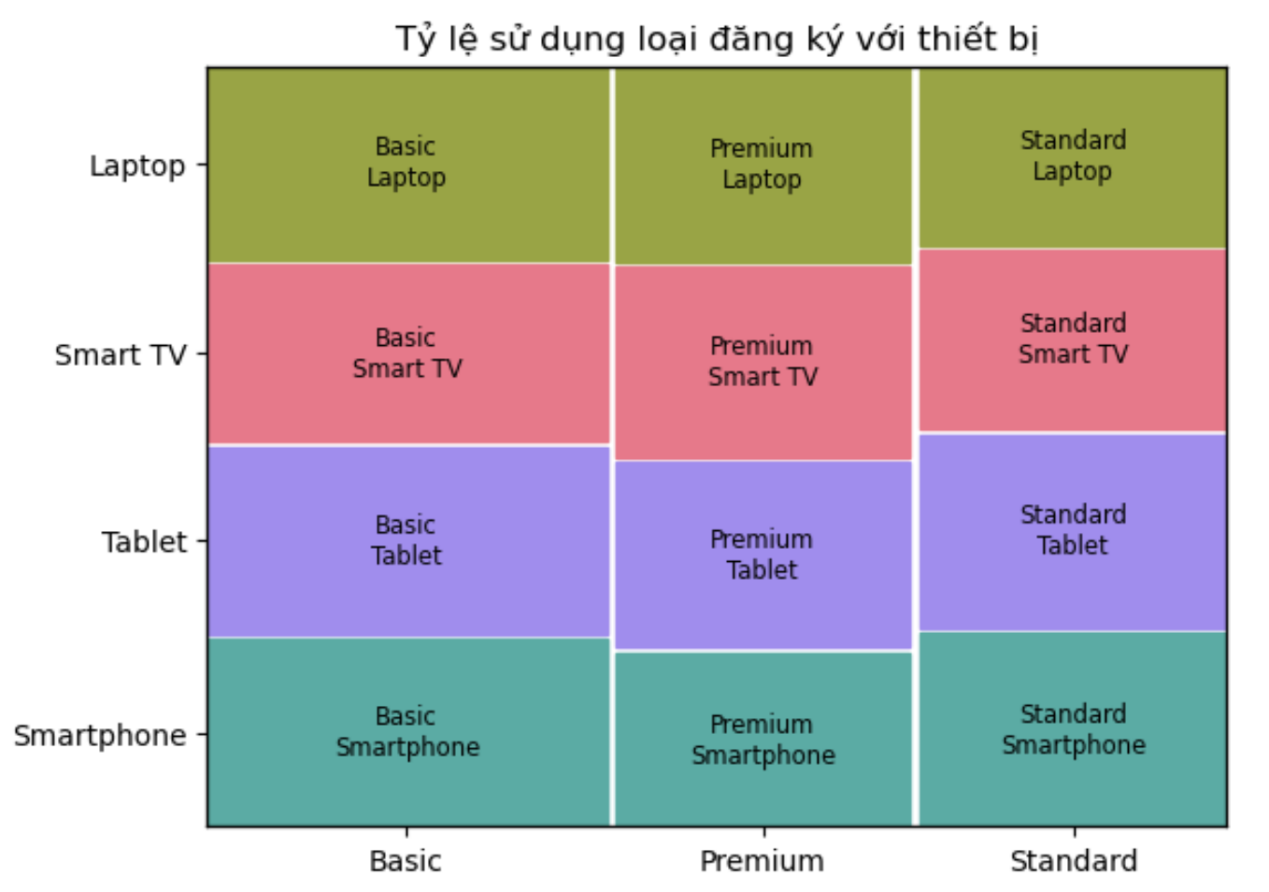
Biểu đồ boxplot: Biểu đồ boxplot cho phép so sánh sự phân phối giữa các nhóm dữ liệu với nhau trên cùng một biểu đồ. Hộp thể hiện giá trị trung vị, tứ phân vị thứ nhất và tứ phân vị thứ 3; thể hiện giá trị tối đa và giá trị tối thiểu.

***- Nhận xét:***

Nhìn chung số lượng người dùng sử dụng trên 4 loại thiết bị (Smartphone, Tablet, Smart TV, Laptop) không có sự chênh lệch nhiều. Thiết bị được người dùng sử dụng cao nhất là Laptop với 636 người dùng. Thiết bị người dùng sử dụng thấp nhất là Smart TV với 610. Smart TV là sự lựa chọn thấp nhất là do dính tiện lợi. Smart TV tuy có màn hình lớn nhưng có nhược điểm khá năng và gần như mua về chỉ để một chỗ không di chuyển. Các thiết bị khác có thể mang đi hàng hàng ngày.

Số lượng doanh thu trên mỗi loại thiết bị là ngang nhau nhưng mức trung bình doanh thu lại có sự chênh lệch với nhau khá lớn. Doanh thu từ người dùng Smartphone tập trung ở mức thấp (lệch dương) là 12 triệu, ba 3 thiết bị còn lại thì doanh thu tập trung ở mức cao (lệch âm) là 13 triệu.

## 2.5. So sánh các thiết bị sử dụng trên từng loại đăng kí



Sử dụng biểu đồ mosaic plot: thể hiện tỷ lệ người dùng sử dụng loại đăng ký trên các thiết bị.

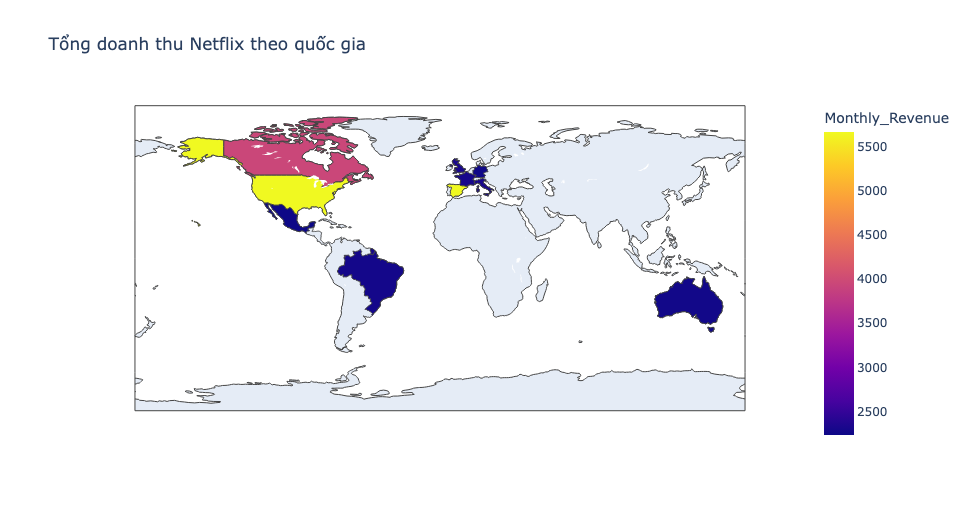
***- Lý do chọn biểu đồ:***

Biểu đồ mosaic: Biểu đồ biểu diễn mối quan hệ giữa các biến phân loại bằng cách phân chia kích thước của các hình chữ nhật hoặc hình vuông theo tỷ lệ với tổng lượng quan sát (có thể sử dụng dữ liệu đa biến).

***- Nhận xét:***

Tỷ lệ sử dụng đăng ký Basic là cao nhất trong 3 loại dịch vụ, loại đăng ký thấp nhất là Premium. Tỷ lệ sử dụng các loại thiết bị của từng loại đăng ký không chênh lệch nhiều. Trong 3 loại đăng ký thì gói Premium có tỷ lệ sử dụng Smartphone thấp nhất nhưng loại có tỷ lệ sử dụng Smart TV nhiều nhất. Điều này có thể thấy rằng gói đăng ký Premium được sử dụng trong hộ gia đình nhiều hơn hai gói còn lại. Tỷ lệ sử dụng Laptop trong gói Standard là thấp nhất so với 2 gói đăng ký còn lại.

## 2.6. Tổng doanh thu theo quốc gia



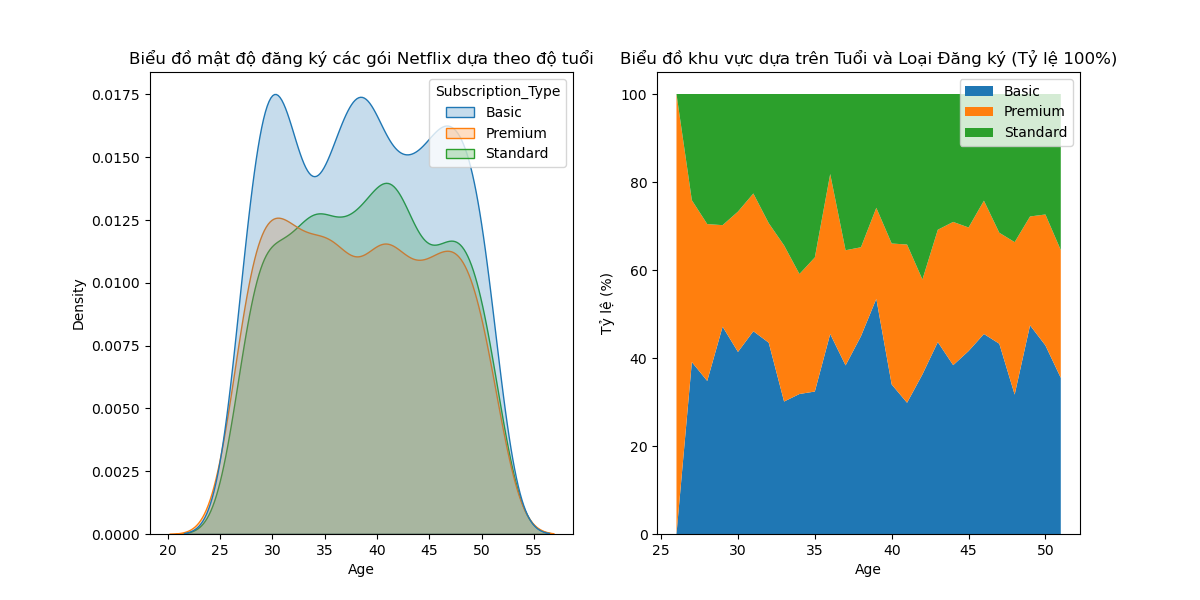
Sử dụng biểu đồ map: Thể hiện vị trí địa lý của từng quốc gia.

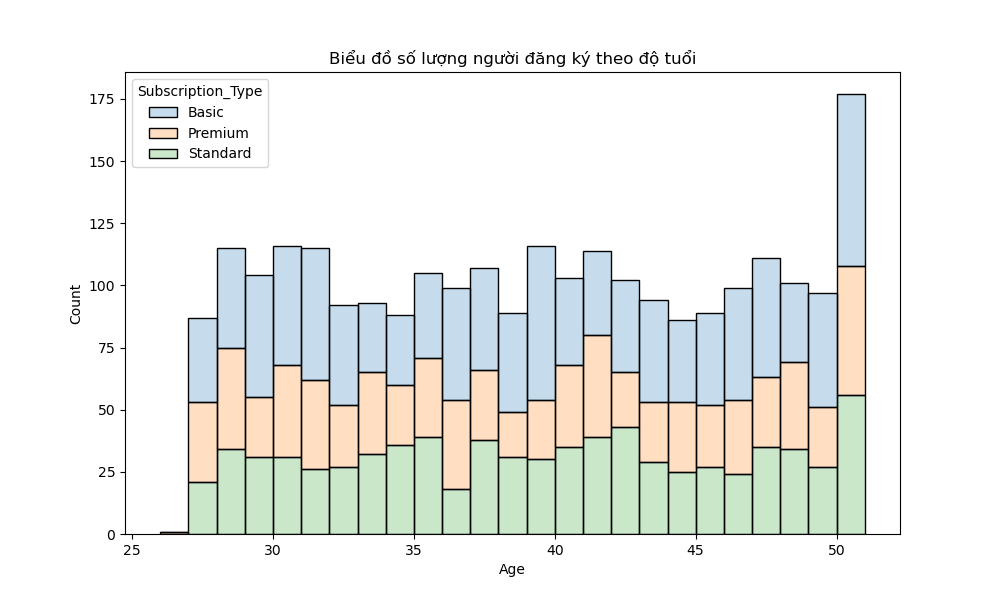
***- Lý do chọn biểu đồ:***

Biều đồ map là một công cụ mạnh mẽ để trực quan hóa dữ liệu địa lý và hiển thị thông tin trên một bản đồ địa lý. Doanh thu thấp hay cao được thể hiện dựa trên màu sắc của từng quốc gia.

***- Nhận xét:*** Nhìn chung , tổng doanh thu của các quốc gia có sự chênh lệch khá lớn. Spain và United States là 2 quốc gia có doanh thu cao nhất lần lượt là 5662 và 5664 (đơn vị doanh thu) thể hiện ở màu vàng theo tháng. Australia, Italy,Brazil, Mexico là các quốc gia có doanh thu thấp nhất dao động từ 2237 đến 2318 (đơn vị doanh thu) thể hiện ở màu xanh nước biển đậm. Còn Canada có tổng doanh thu 3950, so với các nước trên thì Canada ở vị trung bình , thể hiện ở màu hồng.

## 2.7. So sánh số lượng người dùng sử dụng các loại đăng ký theo độ tuổi





Biểu đồ kdeplot: Biểu đồ dùng để thể hiện mật độ đăng ký các gói theo độ tuổi

Biểu đồ stackplot: Biểu đồ thể hiện tỷ lệ đăng ký các gói của Netflix theo độ tuổi

Biểu đồ histplot: Biểu đồ thể hiện tổng số lượng các các gói đăng ký theo độ tuổi đồng thời thể hiện tỷ lệ đăng ký các gói

***- Lý do chọn biểu đồ:***

Kdeplot cho phép nhìn vào phân phối của dữ liệu một cách trực quan hơn so với histogram và được dùng để hiển thị phân bố xác suất của dữ liệu . Điều này giúp dễ dàng nhận ra các đặc điểm như đối xứng, độ nhọn, và đuôi của phân phối.

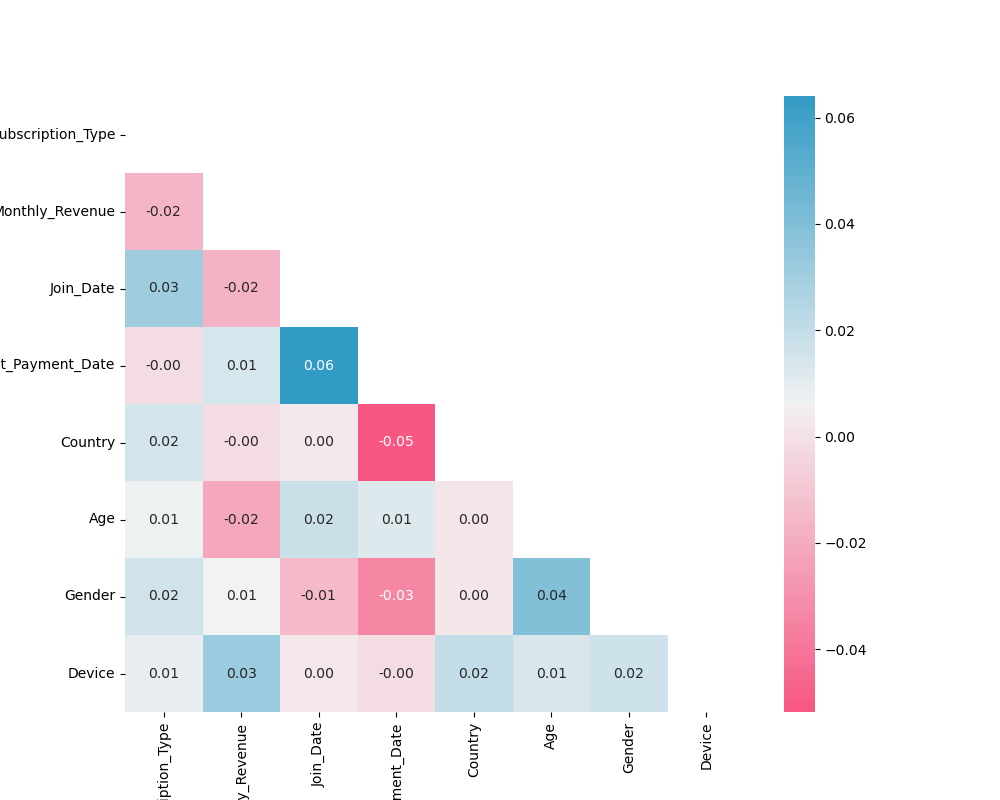
Stackplot được sử dụng để hiển thị sự biến đổi theo tỷ lệ của các thành phần riêng lẻ trong một tổng thể qua thời gian hoặc theo các biến số khác.

Histplot Kết hợp Histogram và KDE (Kernel Density Estimation): Hiển thị đồng thời cả phân bố tần suất (Histogram) và mật độ phân bố (KDE) của dữ liệu dưới dạng cột xếp chồng trên cùng một biểu đồ.

***- Nhận xét:***

Số lượng người sử dụng gói Basic ở mọi lứa tuổi luôn cao hơn số lượng sử dụng 2 gói còn lại. Tỷ lệ người dùng sử dụng các gói đăng ký premium hầu như là không thay đổi theo độ tuổi. Ở những độ tuổi mà số người đăng ký gói Basic tăng thì số người đăng ký gói Standard lại giảm và ngược lại. Ở độ tuổi 27 thì chỉ có người dùng đăng ký gói premium. Số lượng người đăng ký gói Standard cao nhất là ở tuổi 40, ở gói Basic thì ở độ tuổi 38, 38, 50. Mật độ người dùng sử dụng cao nhất của các gói nằm trong độ tuổi khoảng từ 27 đến 50.

## 2.8. Tương quan giữa các thuộc tính



Biểu đồ heatmap: Biểu đồ thể hiện sự tương quan giữa các thuộc tính của dữ liệu

***- Lý do chọn biểu đồ:***

Heatmap là công cụ mạnh mẽ để hiển thị dữ liệu theo một ma trận màu sắc, trong đó màu sắc của từng ô đại diện cho giá trị của một điểm dữ liệu.

Màu xanh càng đậm thể hiện sự đồng biến, ngược lại màu hồng càng đậm thể hiện sự nghịch biến

***- Nhận xét:***

Nhìn chung sự phụ thuộc giữa các thuộc tính trong bộ dữ liệu không cao. Sự tương quan cao nhất là giữa ngày đăng ký và ngày thực hiện thanh toán gần nhất, 2 đại lượng này tỷ lệ thuận với nhau. Có nghĩa là số lượng người đăng ký ở ngày nào tăng thì số lượng người thanh toán ở ngày đó cũng tăng và ngược lại. Ở độ tuổi thì có tỷ lệ nghịch với doanh thu từ độ tuổi đó với tỷ lệ là (0.02). Điều này có nghĩa là ở độ tuổi càng cao thì doanh thu càng giảm. Vì vậy nên tập trung vào những dịch vụ ở độ tuổi trẻ hơn.