Phân tích Dữ liệu kinh doanh giày thể thao kết hợp với Học Máy: Phương pháp phân tích dựa trên Hồi quy tuyến tính, Mạng Học sâu tích chập Và K-means.

1st Đào Việt Anh
Viện Khoa học và Công nghệ Việt Nam - Hàn Quốc
Đại học Công nghệ - ĐHQG Hà Nội
Hà Nội, Việt Nam
2202502@vnu.edu.vn
daovietanh19xx@gmail.com

Tóm tắt nội dung*—Phân tích dữ liệu kinh doanh đã trở thành một vấn đề quan trọng trong các doanh nghiệp và thị trường do xu hướng phát triển của ngành khoa học dữ liệu. Trong đó thị trường bán lại cũng phát triển mạnh mẽ. Hằng năm, một lượng dữ liệu khổng lồ được tạo ra do các hoạt động mua bán sôi nổi trên Internet khiến nó trở thành một tài nguyên quý giá cần được khai phá. Nhờ tính linh hoạt và hiệu quả, các phương pháp học máy được sử dụng rộng rãi để phân tích dữ liệu kinh doanh. Tuy nhiên, để trích lọc các thông tin sâu sắc từ dữ liệu đòi hỏi sự kết hợp của nhiều kỹ thuật khiến các phương pháp học máy đối mặt với nhiều thách thức trong thực tế. Trong khuôn khổ bài báo này, chúng tôi đề xuất một phương pháp phân tích dữ liệu giày thể thao tổng hợp từ trang thương mại điện tử bán lại StockX. Bài báo sẽ tập trung vào việc trả lời các câu hỏi: Thời điểm tốt/xấu nhất để bán giày thể thao? Nơi nào đem lại doanh thu cao nhất, nơi nào có số lượng khách hàng sẵn sàng mua giày thể thao cao nhất? Loại giày thể thao nào đem lại tỷ suất lợi nhuận cao nhất? Thời điểm thích hợp để mua/bán giày thể thao? và Xu hướng màu sắc cũng như loại giày thể thao? Trước tiên, chúng tôi thực hiện chuẩn bị và lọc sạch dữ liệu từ StockX. Sau đó dữ liệu này được đưa qua bước khai phá dữ liệu đặc biệt, được gọi là phân tích dữ liệu thăm dò hay EDA (Exploratory Data Analysis) cho phép khám phá các thông tin, nội dung quan trọng của dữ liệu để có thể xây dựng đặc trưng cho các mô hình. Trong giai đoạn phân loại, chúng tôi sử dụng mô hình học sâu để tạo ra vector đặc trưng cho từng hình ảnh giày và đưa qua thuật toán K-means để phân loại giày nhằm đưa ra các đánh giá về xu hướng màu sắc và xu hướng của giày. Cuối cùng, một mô hình Hồi quy tuyến tính (Linear Regression) được sử dụng để dự đoán giá của một loại giày thể thao thông qua vector đặc trưng trong một thời điểm cho trước. Để xác minh tính khả thi của phương pháp đề xuất, chúng tôi đánh giá mô hình trên hai tập dữ liệu từ StockX 2019 Contest và dữ liệu 22 năm về giày thể thao được khai thác từ StockX.

***Index Terms*—Phân tích dữ liệu kinh doanh, Phân tích dữ liệu thăm dò (EDA), Hồi quy tuyến tính (Linear Regression), Mạng học sâu tích chập, K-means.**

I. GIỚI THIỆU

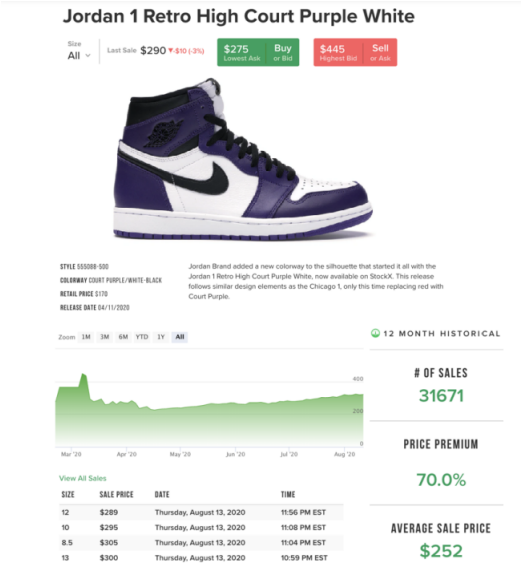
Thị trường bán lại trực tuyến đang phát triển mạnh mẽ. Một trong những ngành phát triển nhanh nhất là thời trang. StockX.com là một trong những thị trường bán lại giày dép trực tuyến. Nó tạo điều kiện cho việc đấu giá giữa người bán và người mua bằng cách xác minh hàng hóa của người bán và chuyển giao cho người mua quốc tế. Hiện nay được coi là một ngành kinh doanh hàng tỷ đô la, trang web này cung cấp thông tin dữ liệu về xu hướng thời trang đường phố thông qua lịch sử giá như một sàn giao dịch chứng khoán. Để mô tả các xu hướng dài hạn dựa trên dữ liệu Web, chúng tôi đã thu thập thông tin về khoảng 23.492 đôi giày dép từ trang web này (Hình 1). Ngoài ra, chúng tôi cũng sử dụng dữ liệu gần 100000 giao dịch bán giày với 50 mẫu giày khác nhau từ StockX 2019 Data Contest cho việc huấn luyện và đánh giá mô hình hồi quy tuyến tính. Các hình ảnh sản phẩm được sử dụng cho một mô hình mạng học sâu tích chập để trích xuất một vector nhúng sau đó sử dụng mô hình K-means để phân loại các vector nhúng qua đó giúp giải thích các đặc điểm thiết kế đặc trưng các thuộc tính màu sắc và hình dáng.

Vector nhúng của phương pháp đề xuất được tạo ra bằng mô hình mạng học sâu tích chập Resnet-18. Giống như các mô hình phân loại và tìm kiếm ảnh, mô hình của chúng tôi tạo ra một vector nhúng 512 chiều trên một không gian ẩn (latent space). Mục tiêu của mô hình là sẽ cố gắng cực đại hóa khoảng cách 2 vector nhúng của 2 ảnh biểu thị hai đối tượng khác nhau và cực tiểu hóa khoảng cách 2 vector nhúng khi 2 đối tượng giống nhau. Thay vì việc đa tăng cường cho từng đặc điểm của ảnh như thay đổi màu sắc hay hình dạng để tạo ra các module cho từng thuộc tính tương ứng thì chúng tôi mô hình hóa toàn bộ các thuộc tính nhằm nắm bắt được nhiều hơn các thuộc tính ẩn ngoài màu sắc và hình dạng. Đột phá kỹ thuật của phương pháp trong bài báo là trích xuất toàn bộ các đặc điểm thị giác của hình ảnh giày dép. Phương pháp này có thể áp dụng cho các mặt hàng thời trang khác.

Nghiên cứu này đem lại một số ý nghĩa. Một số dịch vụ tư vấn thời trang mới dựa trên trí tuệ nhân tạo có thể hỗ trợ ngành công nghiệp trong việc quản lý dữ liệu, dự đoán xu hướng phát triển và đề xuất những thiết kế có thể bền vững hơn [1]. Mô hình học sâu tiên tiến cũng có thể được sử dụng để quan sát các mô hình thời gian của các đối tượng con người

khác và các sản phẩm văn hóa ngoài lĩnh vực thời trang, vì nó không đòi hỏi bất kỳ dữ liệu siêu dữ liệu nào ngoài hình ảnh. Điều này đảm bảo việc học không phụ thuộc vào lĩnh vực và giúp phân tích dữ liệu có lịch sử dài hạn.

Chi tiết triển khai của mô hình và mã nguồn đã được công khai tại https://github.com/daovietanh190499/sneakers_analytic. Vui lòng xem Phụ lục về mô tả kết quả, dữ liệu và phương pháp trích xuất đặc trưng.



Hình 1. Dữ liệu thu thập từ StockX và StockX 2019 Data Contest

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong phần này, chúng tôi giới thiệu một số nghiên cứu về lĩnh lực phân tích dữ liệu kinh doanh và xu hướng thời trang và phương pháp mạng học sâu tích chập tạo vector nhúng cho ảnh.

A. Phân tích xu hướng thời trang

Xu hướng thời trang là các hiện tượng văn hóa và xã hội đã trải qua sự thay đổi liên tục. Trong quá khứ, các nhà khoa học xã hội đã cố gắng nghiên cứu các nguyên tắc cơ bản của xu hướng này. Một số xu hướng có tính mùa và ngắn hạn, trong khi những xu hướng khác có tính lâu dài hơn. Nó cũng có thể theo chu kỳ với những lần hồi sinh định kỳ mang lại sức sống mới hoặc có thể được tái sử dụng như hàng cổ. Cơ chế đằng sau các xu hướng thời trang là đa diện: chúng có thể nội tại trong quá trình tiến hóa hình dạng của một đối tượng [2]–[4] và kết quả của động lực xã hội bên ngoài, như sự khác biệt và phân biệt của con người trong một nhóm. Nhà xã hội học George Simmel nhận thấy rằng thời trang là kết quả của sự căng thẳng giữa mong muốn của cá nhân phù hợp với các xu hướng chủ đạo trong nhóm của mình và mong muốn của anh ta để trở nên độc đáo và nổi bật so với đám đông [5]. Sau đó, Pierre Bourdieu xác định sức mạnh của sự khác biệt là quá trình giải thích cách mọi người trong một nhóm xã hội tuân theo các xu hướng vị thế cụ thể, cho phép họ tách biệt với khẩu vị của một nhóm xã hội khác mà họ muốn tách biệt [6].

Các cơ chế này một phần giải thích cách một số người am hiểu và những người tạo xu hướng đầu tiên xác định những người "tiếp nhận sớm" của một xu hướng trở nên phổ biến và phổ thông theo thời gian. Khi nhiều người tuân theo xu hướng, nó trở nên ngày càng thời thượng đối với những người tạo xu hướng, sau đó họ bỏ nó để chuyển sang điều gì mới bằng cách tạo ra một chu kỳ mới. Người tạo xu hướng sau đó có thể tái định giá lại một xu hướng mà đa số đã bỏ quên trong một thời gian dài. Điều này làm cho thị trường bán lại trở thành một ngành công nghiệp khả thi. Hiệu ứng "hipster" mô tả sự tìm kiếm sự độc quyền trong các xu hướng, đặc biệt là bởi những người thích phục hồi lại các phong cách của quá khứ, và thực tế là một dân số toàn cầu nhạy bén về xu hướng, trái ngược, theo cuộc tìm kiếm này, dẫn đến một đồng bộ ngược nghĩa và do ngành công nghiệp thúc đẩy xảy ra giữa những người muốn tỏ bày tính độc đáo thông qua thời trang [7].

Nhiều mặt hàng thời trang được thị trường toàn cầu ưa chuộng và người tiêu dùng trao đổi thông tin qua mạng. Động lực của xu hướng rất tương tự như thị trường tài chính, nơi một số ít các nhà môi giới chứng khoán hành động dựa trên thông tin độc quyền, được theo sau bởi đám đông nhà đầu tư, tạo ra sự biến động liên tục của giá trị cổ phiếu. Thế giới tài chính luôn sử dụng các phương pháp hệ thống và số học để nắm bắt những quy trình này trong khi nhận ra khó khăn của việc tạo ra các mô hình dự đoán. Trong thời trang, các mô hình dự đoán còn ít hệ thống hơn, vì chúng chủ yếu dựa trên trực giác chủ quan của các chuyên gia về xu hướng và phong cách. Dự báo thời trang là một lĩnh vực nghiên cứu và ngành công nghiệp với một lịch sử dài và hiệu quả của nó đã được đặt ra để tranh luận [8]. Chỉ trong thời gian gần đây, chúng ta mới chứng kiến sự chuyển đổi sang phân tích số liệu lượng lớn trong dự báo thời trang [14, 15].

Ban đầu, các nghiên cứu này dựa trên các tập dữ liệu nhỏ, chẳng hạn như hình ảnh từ các tuần lễ thời trang cụ thể [21]. Ngày nay, sự phát triển của các phương pháp tính toán để phân tích cơ sở dữ liệu lớn cung cấp khả năng thực hiện quan sát bằng cách sử dụng dữ liệu trên web và học máy. Các nhà nghiên cứu trong lĩnh vực khoa học dữ liệu, khoa học xã hội và nhân văn số đã công bố nhiều nghiên cứu sử dụng phương pháp tính toán để phân tích các thay đổi về phong cách, hình dạng và nội dung trong văn học, nghệ thuật hình ảnh, âm nhạc phổ biến và truyền thông đại chúng trong một khoảng thời gian dài [1, 20]. Trong lĩnh vực thời trang, hình ảnh đã được phân tích tự động, ví dụ như phân tích tự động của các mặt hàng quần áo [45], nhận dạng sản phẩm và phân loại phong cách [22, 25]. Phân tích trên các cơ sở dữ liệu lớn cho phép chúng ta khám phá các xu hướng tổng quát rộng lớn, chẳng hạn như sự thay đổi mùa trong màu sắc quần áo [16, 42], và phát hiện và dự đoán các thay đổi về phong cách thời trang qua không gian và thời gian mà tránh được sự quan sát trực tiếp của con người [28, 30].

B. Các phương pháp tạo vector nhúng

Sự thành công của mạng học sâu tích chập trong việc biểu diễn đặc trưng đã giúp nó trở thành một kỹ thuật tiêu chuẩn trong việc truy xuất hình ảnh. Các mô hình được huấn luyện

trước trên các tập dữ liệu phổ biến như ImageNet (Deng et al. 2009), Landmarks (Babenko et al. 2014), v.v. có thể được sử dụng để trích xuất các đặc trưng của hình ảnh. Đặc biệt, các lớp tích chập đã được chứng minh là có lợi nhất trong việc truy xuất hình ảnh (Babenko et al. 2014; Radenovic, Tolias, và Chum 2016; Gordo et al. 2016; ´ Razavian et al. 2016). Sau đó, tìm kiếm hàng xóm gần nhất được sử dụng trên các vector đặc trưng để tìm các hình ảnh tương tự nhất với một truy vấn. Các bộ dữ liệu thường được thu thập từ internet, dẫn đến các hình ảnh của cùng một đối tượng/địa danh được hiển thị ở nhiều góc độ, ánh sáng và điều kiện khác nhau. Sự đa dạng này thường dẫn đến việc hình thành các mặt phẳng trên không gian đặc trưng không thích hợp cho việc xếp hạng dựa trên phép đo khoảng cách. Một số nghiên cứu khác của Sungkyu Park et al. 2022; ‘ tập trung vào việc xác định vector nhúng cho nội dung của bức ảnh như các thuộc tính về màu sắc và hình dáng bằng cách tạo ra các module và đa tăng cường dữ liệu để chúng có thể học một trong các thuộc tính thị giác mong muốn.

III. BÀI TOÁN VÀ DỮ LIỆU

Trong phần này, chúng tôi giới thiệu về các câu hỏi nghiên cứu và các tập dữ liệu.

A. Câu hỏi nghiên cứu

Các câu hỏi nghiên cứu trong khuôn khổ bài báo bao gồm: Đầu tiên, chúng tôi xác định thời điểm tốt/xấu nhất để bán giày thể thao? Việc này giúp các nhà kinh doanh có thể lựa chọn thời điểm nhằm thu lợi nhuận hiệu quả. Thứ hai, xác định vùng đem lại doanh thu cao nhất, vùng có số lượng khách hàng sẵn sàng mua giày thể thao cao nhất cũng như loại giày thể thao nào đem lại tỷ suất lợi nhuận cao nhất? chúng tôi sẽ trả lời câu hỏi này thông qua việc phân tích dữ liệu thăm dò EDA. Cuối cùng, xác định xu hướng màu sắc cũng như loại giày thể thao và dự đoán giá cả từng loại giày? Chúng tôi sẽ sử dụng các phương pháp Học Máy để trích xuất các thông tin giúp các nhà đầu tư và thiết kế chiến lược đầu tư và quyết định các mẫu giầy trong tương lai.

B. Mô tả dữ liệu

Trong bài báo này, chúng tôi sử dụng hai tập dữ liệu từ hai nguồn bao gồm:

- Dữ liệu 22 năm về giày thể thao khai thác từ trang StockX.com dựa trên nghiên cứu của Sungkyu Park et al. 2022;
- Dữ liệu từ StockX 2019 Data Contest trên Kaggle.

Đối với tập dữ liệu 22 năm về giày thể thao khai thác từ trang StockX.com, chúng tôi đã thu thập tất cả các đặc điểm được cung cấp trên trang web, bao gồm giá bán lẻ, lịch sử giá bán lại, hình ảnh sản phẩm, thương hiệu và ngày phát hành. Chúng tôi đã tìm được thông tin cho khoảng 23.492 đôi sneaker từ những năm trước đây (ví dụ, năm phát hành của một số đôi sneaker trở lại năm 1985). Dữ liệu giao dịch giá có sẵn từ năm 2012 kể từ khi nền tảng ra mắt. Hình ảnh sản phẩm có độ phân giải ngang, vì vậy chúng tôi đã thêm các viền màu trắng vào phần trên và dưới để tạo ra hình ảnh vuông. Sau đó, chúng tôi đã thay đổi kích thước hình ảnh thành 256×256 pixel để giảm kích thước cho việc xây dựng đặc trưng và sử dụng mạng đã được huấn luyện trước trên ImageNet (ví dụ, ResNet với 18 lớp). Việc huấn luyện trước trên một tập dữ liệu quy mô lớn đảm bảo rằng mô hình nắm bắt được các đặc điểm hình ảnh có thể tổng quát và giúp tạo ra các chỉ số chất lượng cao [10]. Trước khi phân tích, chúng tôi đã loại bỏ bất kỳ sản phẩm nào không có hình ảnh giày thích hợp, ví dụ như chỉ hiển thị hộp giày. Dữ liệu cuối cùng bao gồm 11,0 gigabyte, bao gồm cả hình ảnh và siêu dữ liệu cho 22.331 đôi sneaker có hình ảnh hợp lệ.

Đối với tập dữ liệu từ StockX 2019 Data Contest trên Kaggle, tập dữ liệu bao gồm một mẫu ngẫu nhiên của tất cả các giao dịch bán hàng Off-White x Nike và Yeezy 350 từ ngày 1/9/2017 (tháng mà Off-White ra mắt bộ sưu tập "The Ten") cho đến hiện tại. Tổng cộng có 99.956 giao dịch trong tập dữ liệu này, trong đó có 27.794 giao dịch Off-White và 72.162 giao dịch Yeezy. Mẫu được lựa chọn từ các giao dịch tại Hoa Kỳ. Giống với tập dữ liệu 22 năm về giày thể thao, chúng tôi tiến hành lọc sạch các bản ghi bị thiếu trường dữ liệu và phân bố phần còn lại vào các thuộc tính chung trong một bảng.

IV. PHƯƠNG PHÁP ĐỀ XUẤT

A. Phân tích dữ liệu thăm dò - Exploratory Data Analysis

EDA (Exploratory Data Analysis) là quá trình nghiên cứu và khám phá dữ liệu để hiểu và tìm ra thông tin hữu ích từ tập dữ liệu. EDA giúp chúng ta xác định các đặc điểm quan trọng, mối quan hệ giữa các biến và các mẫu dữ liệu, cũng như phát hiện các điểm ngoại lệ và thiếu sót trong dữ liệu. Đây là giai đoạn quan trọng trong quy trình phân tích dữ liệu và thường được thực hiện trước khi áp dụng các phương pháp phân tích tiên tiến hơn.

Trong nghiên cứu này, chúng tôi thực hiện EDA trên các tập dữ liệu theo các bước sau:

a) Xem xét toàn bộ dữ liệu: Điều này bao gồm xem xét số lượng mẫu và các biến, kiểu dữ liệu của từng biến và đảm bảo rằng chúng đúng định dạng mà chúng tôi mong đợi. Chúng tôi giữ nguyên các trường của tập dữ liệu bao gồm: ’Order Date’, ’Brand’, ’Sneaker Name’, ’Sale Price’, ’Retail Price’, ’Release Date’, ’Shoe Size’, ’Buyer Region’. Sau đó, thêm các trường mới bao gồm: ’Bought for Retail’, ’Bought for Less Than Retail’, ’Bought for More Than Retail’, ’Resale per Thousand per Region’, ’Resale per Week per Region’, ’Average Retail Per Week’.

b) Trực quan hóa dữ liệu: Chúng tôi sử dụng các đồ thị để trực quan hóa dữ liệu, xem thêm phần phụ lục. Bằng cách này chúng tôi có thể xem xét và đánh giá dữ liệu một cách trực quan. Để làm mối quan hệ và xu hướng trong dữ liệu chúng tôi thực hiện việc biểu diễn các thông tin sau:

- Số lượng giày bán trong một khoảng thời gian
- Số lượng giày bán theo khu vực
- Số lượng giày bán theo thương hiệu
- Số lượng giày bán theo cỡ
- Gía trị trung bình từng loại giày

- Mối tương quan giữa tổng số bán lại trên một khu vực và thời gian tính theo tuần
- Mối tương quan giữa giá bán lẻ, giá bán buôn và mật độ mua hàng tại từng khu vực

c) Phân tích dữ liệu: Chúng tôi đã xem xét quy mô và mô hình thời gian trong dữ liệu mô tả của các đôi giày. Số lượng sản phẩm và số lượng thương hiệu trên thị trường bán lại đã tăng nhanh chóng trong thập kỷ qua, như được biểu thị bởi sự tăng trưởng mạnh mẽ của ngành kinh doanh này. Kể từ mùa xuân năm 2019, giao dịch hàng quý đã đạt đến quy mô hàng triệu trên nền tảng này, và hơn 10.000 mặt hàng đã được bán hàng quý. Chúng tôi xác định lợi nhuận hoặc tỷ lệ bán lại cao như giá bán lại trừ giá bán lẻ cho mỗi giao dịch, trong đó mỗi giá đã được điều chỉnh để tính toán sự lạm phát của đồng USD. Khi có nhiều đôi giày được bán thông qua nền tảng này, tỷ lệ bán lại trên mỗi giao dịch có xu hướng giảm. Một giao dịch điển hình dẫn đến lợi nhuận từ 60 đến 80 đô la Mỹ cho người bán lại, chúng tôi giả định người bán lại đã mua sản phẩm với giá bán lẻ. Tuy nhiên, 10% đôi giày có tỷ lệ bán lại cao nhất dẫn đến một lợi nhuận từ 320 đến 400 đô la Mỹ cho mỗi giao dịch, cho thấy một lợi nhuận đáng kể khi xem xét giá bán lẻ của đôi giày. Tỷ lệ bán lại cũng có xu hướng tăng theo tuổi đời của đôi giày (tức là ngày phát hành).

d) Kiểm tra giả định và rút ra kết luận: Cuối cùng, sau khi đã thực hiện các bước EDA trên tập dữ liệu, ta có thể kiểm tra các giả định và rút ra các kết luận dựa trên các quan sát và phân tích đã được thực hiện. Điều này giúp ta hiểu rõ hơn về dữ liệu và chuẩn bị cho các bước tiếp theo trong quy trình phân tích dữ liệu. Chúng tôi đưa ra một số kết luận sau:

- Thời điểm xấu nhất để bán lại giày là trong vòng 9 tuần sau ngày tung sản phẩm mới ra thị trường do thị trường đã bão hòa sản phẩm mới, một trong các thời điểm khác là từ 52 đến 64 tuần sau ngày ra mắt.
- Bán lại trong khoảng 3-5 tuần trước khi sản phẩm mới được tung ra thị trường là thời điểm tốt nhất.
- Khi thị trường bắt đầu suy giảm, người dùng sẽ sẵn sàng trả một mức giá cao cho các sản phẩm tồn kho.
- Trong khoảng 8-23 ngày và 30-48 ngày từ khi sản phẩm được tung ra thị trường là thời điểm tốt để mua.
- Các khu vực như California và New York là những nơi có thị trường bán lại sôi động nhất, trong khi đó tính theo mật độ là Oregon và New York.

B. Mô hình

Các đặc điểm hình ảnh của thiết kế giày thể thao được học thông qua một mô hình mạng học sâu tích chập. Chúng tôi đã phát triển một mô hình tạo vector nhúng dựa trên Mạng học sâu Resnet-18 bằng cách thay thế lớp kết nối đầy đủ với đầu ra 512 chiều. ResNet-18 là một kiến trúc mạng học sâu (deep learning) phổ biến được sử dụng cho các nhiệm vụ truy xuất ảnh và phân loại ảnh. Đây là một trong những phiên bản của Residual Network (ResNet), một dạng mạng nơ-ron tích chập (CNN) được xây dựng để vượt qua vấn đề sự mất mát thông tin khi mạng trở nên sâu. ResNet-18 bao gồm 18 lớp, bao gồm các lớp tích chập, Lớp chuẩn hóa và lớp kết nối đầy đủ. Mô hình này có một kiến trúc chính với các khối cơ bản được gọi là residual blocks. Mỗi residual block có hai lớp tích chập và một kết nối trực tiếp, cho phép thông tin truyền qua mạng mà không bị ảnh hưởng bởi vấn đề mất mát thông tin. Mạng Resnet-18 được huấn luyện bằng phương pháp học giám sát. Học giám sát giả định một nhiệm vụ cố gắng tập hợp các biểu diễn của các trường hợp tương tự (gọi là mẫu dương) gần nhau trong không gian nhúng, trong khi đẩy xa các biểu diễn của các trường hợp không tương tự (gọi là mẫu âm). Người ta có thể coi các biến thể của cùng một hình ảnh giày thể thao (ví dụ: phóng to, cắt, lật) là các mẫu dương và coi hai hình ảnh giày sneaker riêng biệt là các mẫu âm. Phương pháp này có thể hiệu quả học các đặc trưng không thay đổi đối với việc tăng cường dữ liệu. Mô hình chúng tôi sử dụng hàm mất mát Triplet Loss. Hàm Triplet Loss có công thức như sau:

$$L = \sum_{i}^{N} [||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha]$$

Trong đó:

- i là đầu vào thứ i.
- $f(x_i)$ là các vector nhúng đầu ra.
- $||||_2^2$ là chuẩn hóa l_2
- a là mẫu neo, p là mẫu dương và n là mẫu âm.
- α là hệ số thêm.

C. K-means

K-means là một thuật toán phân cụm không giám sát, ý tưởng chính của nó là nhóm dữ liệu thành các cụm, giảm thiểu khoảng cách giữa các cụm khác nhau và tăng cường khoảng cách trong cùng một cụm. Đầu tiên, K-means chọn ngẫu nhiên điểm dữ liệu làm trọng tâm ban đầu từ tập dữ liệu. Tiếp theo, phân phối dữ liệu còn lại vào các cụm tương ứng theo nguyên tắc của việc tính khoảng cách. Sau đó, tính trung bình của dữ liệu trong mỗi cụm để cập nhật trung tâm của cụm. Sau khi các trung tâm của các cụm được cập nhật, thủ tục phân bổ dữ liệu vào các cụm được lặp lại. Lặp lại quy trình phân bổ này cho đến khi thoả mãn điều kiện sau: a) đạt được số lần lặp cố định; b) các trọng tâm không thay đổi hoặc thay đổi trong ngưỡng giữa các lần lặp lại. Trong nghiên cứu của chúng tôi, chúng tôi sử dụng K-means để chia tập vector nhúng của giày thành 5 nhóm khác nhau và phân tích cho mỗi nhóm sử dụng phân tích dữ liệu thăm dò.

D. Hồi quy tuyến tính

Hồi quy tuyến tính là một phương pháp trong lĩnh vực thống kê và machine learning được sử dụng để dự đoán một biến mục tiêu dựa trên các biến đầu vào. Nó giúp xác định mối quan hệ tuyến tính giữa các biến đầu vào và biến mục tiêu.Trong hồi quy tuyến tính, chúng ta giả định rằng mối quan hệ giữa biến đầu vào và biến mục tiêu có thể được mô tả bằng một hàm tuyến tính. Mục tiêu là tìm ra một mô hình hồi quy tuyến tính tốt nhất để dự đoán giá trị của biến mục tiêu dựa trên các giá trị biến đầu vào.

Mô hình hồi quy tuyến tính có dạng:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Trong đó:

- y là biến mục tiêu cần dự đoán.
- $x_1, x_2, ..., x_n$ là các biến đầu vào.
- $\beta_0, \beta_1, \beta_2, ..., \beta_n$ là các hệ số hồi quy (slope) tương ứng với từng biến đầu vào.
- ϵ là sai số ngẫu nhiên.

Mục tiêu trong hồi quy tuyến tính là tìm ra các giá trị tối ưu của các hệ số hồi quy (β) để mô hình hồi quy tuyến tính phù hợp nhất với dữ liệu. Quá trình này thường được thực hiện bằng cách tìm giá trị nhỏ nhất cho hàm mất mát (loss function) như bình phương sai (mean squared error) hoặc tương tự. Khi mô hình hồi quy tuyến tính được xác định, chúng ta có thể sử dụng nó để dự đoán giá trị mới cho biến mục tiêu dựa trên giá trị biến đầu vào. Đồng thời, các hệ số hồi quy cũng có thể được sử dụng để đánh giá mức độ tác động của từng biến đầu vào lên biến mục tiêu. Trong bài báo này, chúng tôi sử dụng Hồi quy tuyến tính để dự đoán giá trị của giày trong một thời điểm. Chúng tôi sử dụng vector nhúng, thời điểm ra mắt, thời điểm bán, kích cỡ và vị trí địa lý làm giá trị đầu vào và loại bỏ các thuộc tính khác bao gồm tên thương hiệu và tên sản phẩm và sử dụng giá bán lẻ làm đầu ra. Mô hình được huấn luyện và đánh giá trên tập StockX 2019 Data Contest

V. THÍ NGHIỆM VÀ KẾT QUẢ

Trên StockX.com, giày thể thao được gán nhãn thuộc một trong 16 danh mục sản phẩm, trong đó chúng tôi chọn 8 lớp hàng đầu theo tần suất. Các danh mục này bao gồm Adidas (4.123 sản phẩm), Air Jordan (3.675 sản phẩm), Air Max (2.839 sản phẩm), Nike Basketball (1.256 sản phẩm), Air Force (1.213 sản phẩm), Nike SB (816 sản phẩm), LeBron (657 sản phẩm) và Kobe (428 sản phẩm). Nền tảng bán lại này cũng gán cho mỗi đôi thể thao bảy loại khách hàng mục tiêu, trong đó chúng tôi chọn 5 loại hàng đầu theo tần suất: nam (3.000 sản phẩm), nữ (2.344 sản phẩm), trẻ em (1.460 sản phẩm), trẻ mẫu giáo (420 sản phẩm) và trẻ nhỏ (340 sản phẩm). Vì các lớp không đồng đều, chúng tôi chọn một số sản phẩm tương tự cho mỗi lớp.

Đối với tập StockX 2019 Data Contest, dữ liệu được chia làm 2 phần theo tỷ lệ 80-20 dùng để huấn luyện mô hình Hồi quy tuyến tính. Số lượng nhãn cũng được chia đều theo 2 tập.

Chúng tôi tiến hành thực hiện các bước EDA, huấn luyện mô hình và phân loại sử dụng K-means và thu được kết quả như các hình dưới.

VI. KẾT LUẬN

Trong bài báo này, chúng tôi đề xuất một phương pháp phân tích dữ liệu kinh doanh dựa trên EAD và Học Máy. Chúng tôi kết hợp sử dụng thông tin về màu sắc và hình dạng để nhúng các thiết kế giày thể thao từ một tập dữ liệu của hình ảnh trên Web. Kết quả phân tích dữ liệu của chúng tôi đã tiết lộ các mẫu hội tụ và độc đáo trong thiết kế của các nhà thiết kế giày thể thao hàng đầu trong hai thập kỷ cùng với đó là các thông tin hữu ích cho chiến lược bán hàng và quảng bá.

Phương pháp và kết quả được trình bày trong bài báo này có ý nghĩa đối với việc dự đoán xu hướng thời trang một cách tốt hơn. Phương pháp của chúng tôi có thể hỗ trợ các nỗ lực trong việc hiểu sự tiến hóa phong cách trong thời trang đại trà bằng các phương pháp định tính hoặc không sử dụng mạng neural. Các phương pháp khoa học dữ liệu có sử dụng thực tế trong việc giúp các nhà thiết kế nắm bắt những đặc điểm quan trọng của xu hướng thời trang từ dữ liệu lớn mà khó tìm thấy bằng cách thủ công. Trong khi việc sử dụng trí tuệ nhân tạo trong ngành công nghiệp thời trang tập trung vào việc tăng doanh số bán hàng và tự động hóa quy trình, nghiên cứu của chúng tôi là một nỗ lực đầu tiên trong việc sử dụng trí tuệ nhân tạo để nắm bắt xu hướng thiết kế. Hướng đi của các sản phẩm cao cấp có thể cung cấp thông tin cho nhà thiết kế trong việc nhận thức về nhận định của người dùng về hàng hóa có tính sưu tầm. Hơn nữa, khách hàng có thể sử dụng thông tin này để dự đoán xu hướng mới và lập kế hoạch chiến lược đầu tư cho việc sưu tầm.

TÀI LIỆU

[1] Amaia Salvador, Michal Drozdzal, Xavier Giro-i Nieto, and Adriana Romero. 2019. Inverse cooking: Recipe generation from food images. In proc. of the CVPR. 10453–10462.

[2] George Kubler. 1962. The shape of time : remarks on the history of things. Yale University Press.

[3] Colin Martindale. 1990. The clockwork muse: The predictability of artistic change. Basic Books.

[4] Martin Siefkes and Emanuele Arielli. 2018. The Aesthetics and Multimodality of Style. Peter Lang

[5] Georg Simmel. 1957. Fashion. Amer. J. Sociology 62, 6 (1957), 541–558

[6] Pierre Bourdieu. 1984. Distinction. Routledge.

[7] Jonathan D Touboul. 2019. The hipster effect: When anti-conformists all look the same. Discrete Continuous Dynamical Systems-B 24, 8 (2019), 4379–4415.

[8] Regina Lee Blaszczyk and Ben (eds.) Wubs. 2018. The Fashion Forecasters. A Hidden History of Color and Trend Prediction. Bloomsbury.

[9] Mikayla DuBreuil and Sheng Lu. 2020. Traditional vs. big-data fashion trend forecasting: an examination using WGSN and EDITED. International Journal of Fashion Design, Technology and Education 13, 1 (2020), 68–77.

[10] Takao Furukawa, Chikako Miura, Mori Kaoru, Sou Uchida, and Makoto Hasegawa. 2019. Visualisation for analysing evolutionary dynamics of fashion trends. International Journal of Fashion Design, Technology and Education 12 (2019), 1–13.

[11] Shintami Chusnul Hidayati, Kai-lung Hua, Wen-Huang Cheng, and Shih-Wei Sun. 2014. What are the Fashion Trends in New York?. In proc. of the ACM Multimedia.

[12] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. 2017. Fashion forward: Forecasting visual style in fashion. In proc. of the ICCV. 388–397

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. Advances in Neural Information Processing Systems (2020).

[14] Wei Yang, Ping Luo, and Liang Lin. 2014. Clothing co-parsing by joint image segmentation and labeling. In proc. of the CVPR. 3182–3189.

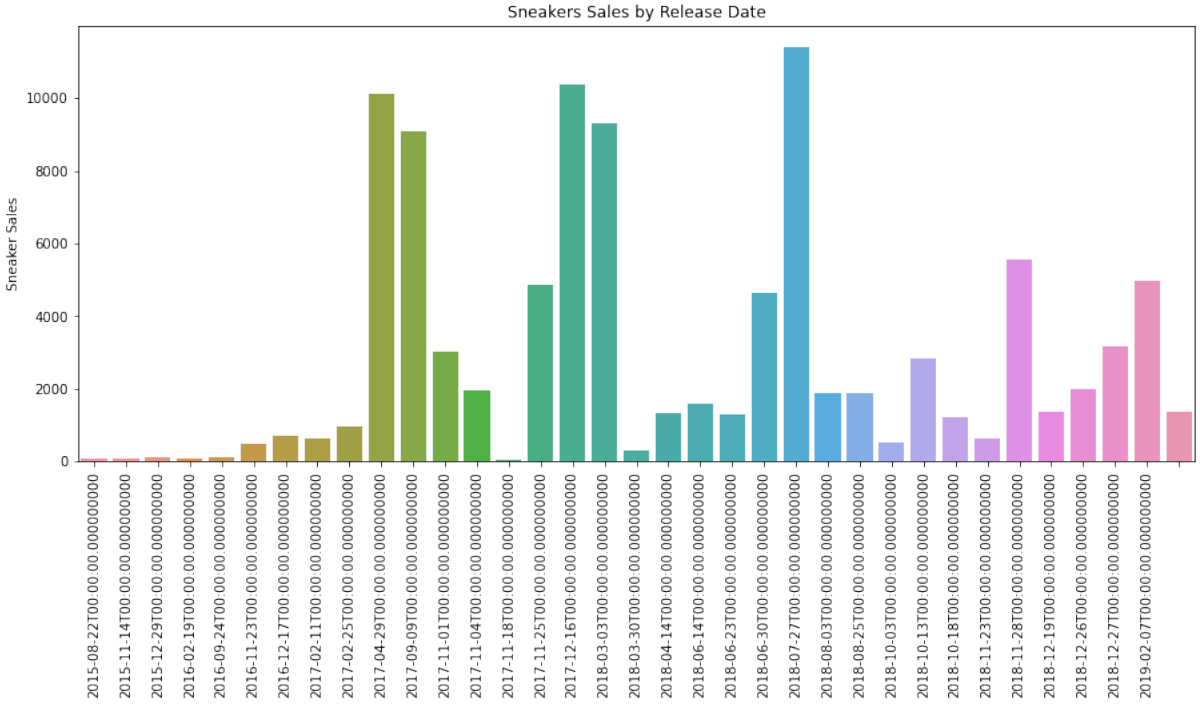
[15] Youngseung Jeon, Seungwan Jin, and Kyungsik Han. 2021. FANCY: Humancentered, Deep Learning-based Framework for Fashion Style Analysis. (2021), 2367–2378.

[16] M. Hadi Kiapour, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. 2014. Hipster Wars: Discovering Elements of Fashion Styles. In proc. of the ECCV. 472–488.

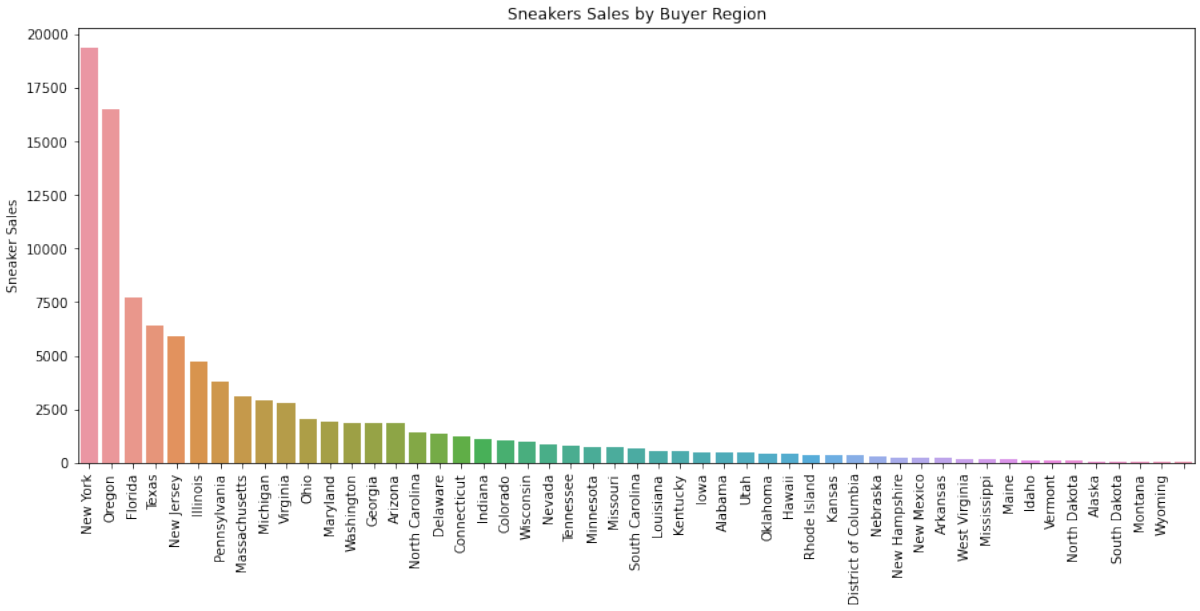
[17] Ahyoung Han, Jihoon Kim, and Jaehong Ahn. 2021. Color Trend Analysis using Machine Learning with Fashion Collection Images. Clothing and Textiles Research Journal (2021).

[18] Sirion Vittayakorn, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. 2015. Runway to Realway: Visual Analysis of Fashion. In proc. of the IEEE Winter Conference on Applications of Computer Vision. 951–958.

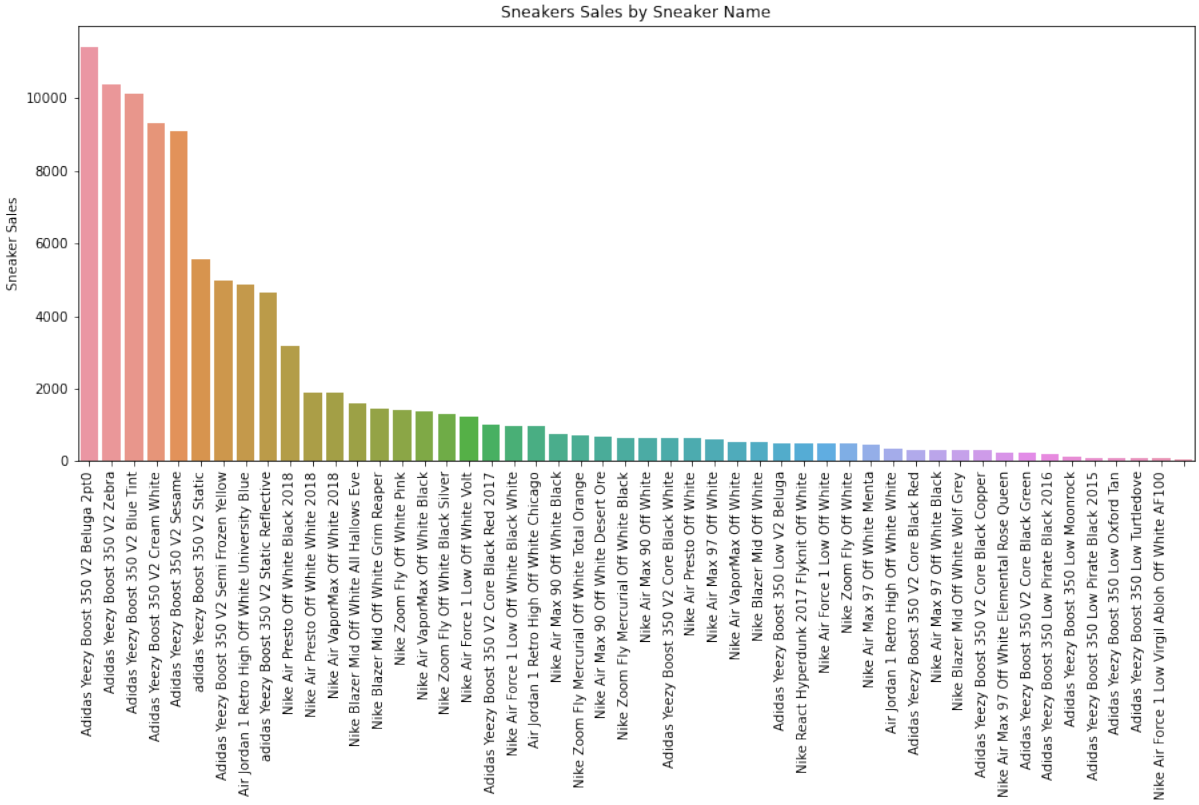
[19]] Utkarsh Mall Mall, Kevin Matzen, Bharath Hariharan, Noah Snavely, and Kavita Bala. 2019. GeoStyle: Discovering Fashion Trends and Events. arXiv:1908.11412 [cs.CV]



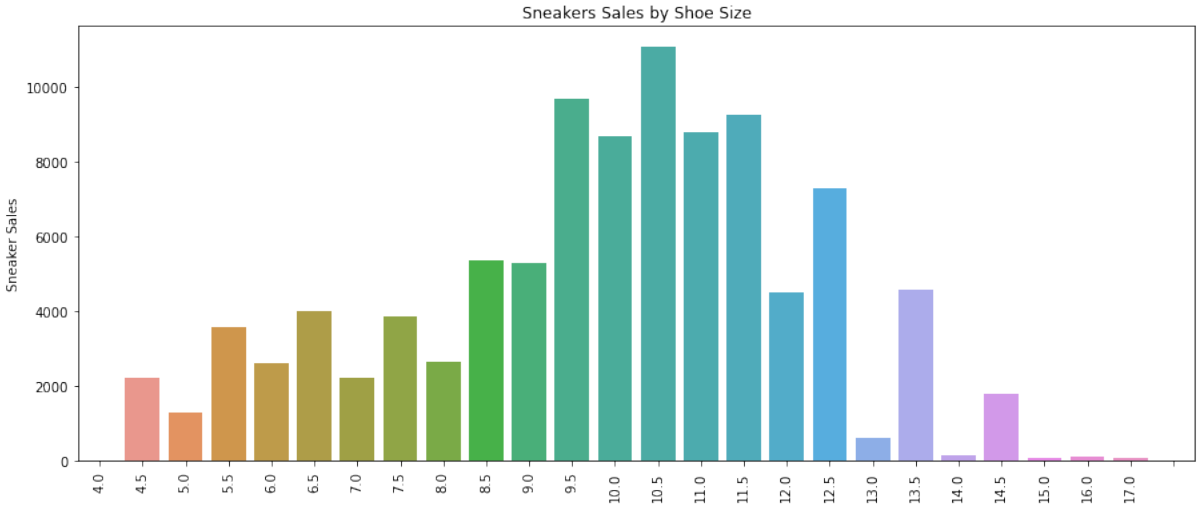
Hình 2. Số lượng giày bán trong một khoảng thời gian



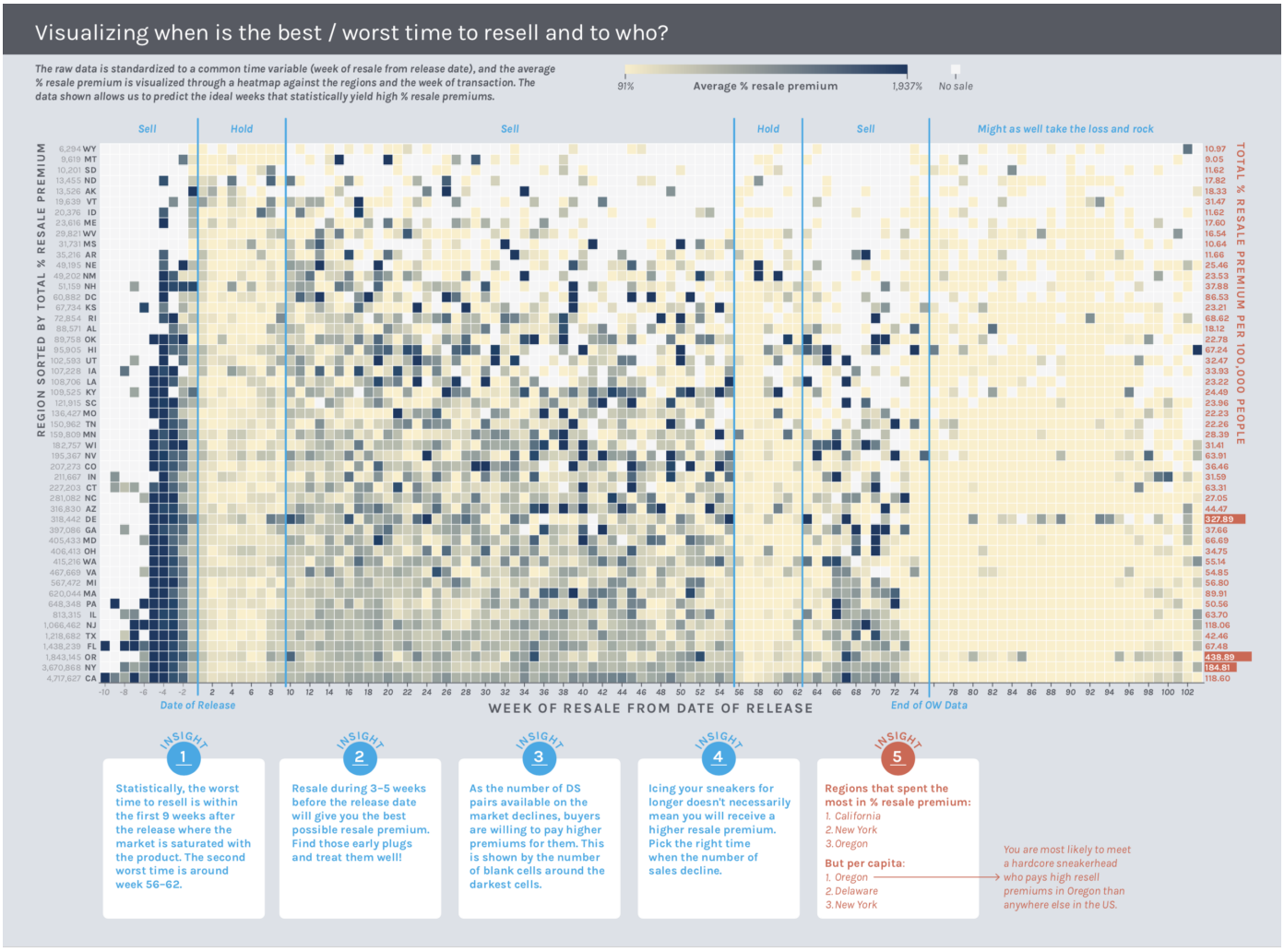
Hình 3. Số lượng giày bán theo khu vực



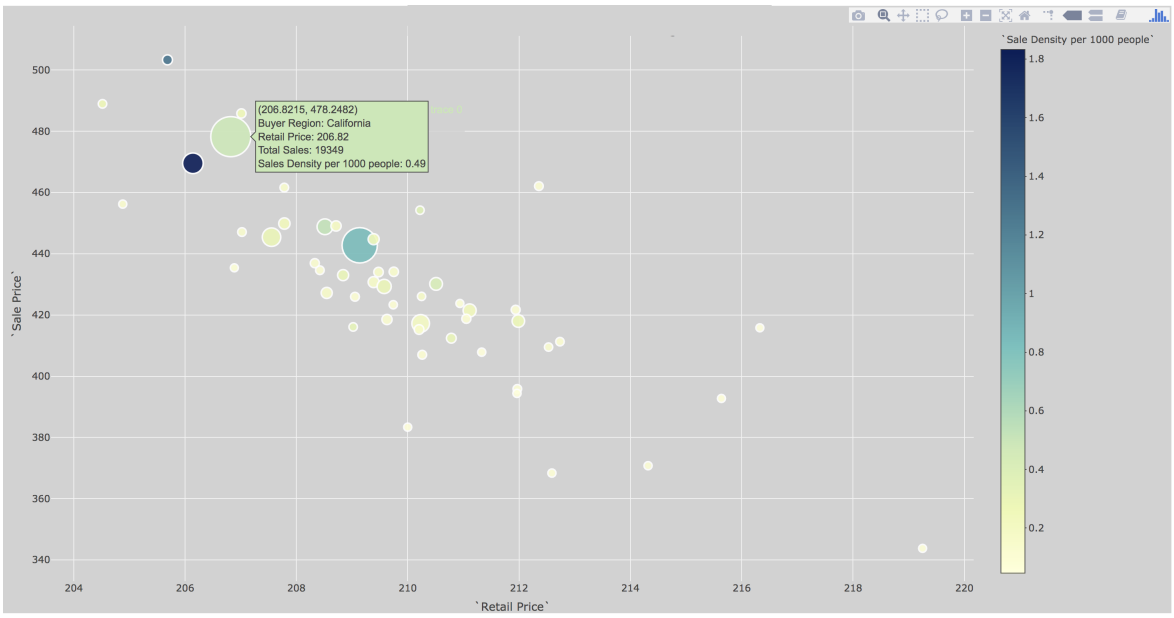
Hình 4. Số lượng giày bán theo thương hiệu



Hình 5. Số lượng giày bán theo cỡ



Hình 6. Mối tương quan giữa tổng số bán lại trên một khu vực và thời gian tính theo tuần



Hình 7. Mối tương quan giữa giá bán lẻ, giá bán buôn và mật độ mua hàng tại từng khu vực

[20] Kevin Matzen, Kavita Bala, and Noah Snavely. 2017. StreetStyle: Exploring world-wide clothing styles from millions of photos. arXiv:1706.01869 [cs.CV]