BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo: Tương tác dữ liệu trực quan

TRỰC QUAN, PHÂN TÍCH VÀ DỰ ĐOÁN DOANH SỐ BÁN HÀNG CỦA MỘT SIÊU THỊ TẠI THỊ TRƯỜNG MỸ

Giảng viên : Ths. Lê Quang Thái

Nhóm sinh viên thực hiện : Nhóm 3

Văn Mai Thanh Nhật	20133076
Huỳnh Minh Phước	20133082
Trần Nguyên Hạnh	20133013
Trần Đông	20133035

TP. Hồ Chí Minh, tháng 5 năm 2023

Mục lục

PHẦN 1 – TÓM TẮT4

PHẦN 2 – GIỚI THIỆU5

PHẦN 3 – DỮ LIỆU7

3.1 Tiền xử lý8

PHẦN 4 – TRỰC QUAN HOÁ DỮ LIỆU10

4.1 Khách hàng:10

4.1.1. Những khách hàng có doanh số cao nhất10

4.1.2 Tỷ lệ doanh thu theo từng phân khúc khách hàng10

4.2 Địa chỉ khách hàng:11

4.2.1. Những tiểu bang có doanh số cao nhất11

4.2.2. Những thành phố có doanh số cao nhất12

4.2.3. Tỷ lệ doanh thu theo từng vùng miền12

4.2.4. Tỷ lệ doanh thu theo từng phương thức vận chuyển13

4.3 Sản phẩm:13

4.3.1. Những sản phẩm có doanh số cao nhất13

4.3.2. Tỷ lệ doanh thu theo từng danh mục sản phẩm, danh mục phụ sản phẩm14

4.4 Doanh số bán được theo từng năm của từng loại danh mục sản phẩm:15

4.5 Doanh số bán được của từng bang trên bản đồ nước Mỹ :15

PHẦN 5 - MÔ HÌNH HÓA DỮ LIỆU17

5.1 Phân tích đa biến dùng K-Means Clustering (Phân cụm K-means)17

5.2 Mô hình dự đoán chuỗi thời gian18

5.2.1 Mô hình dự đoán chuỗi thời gian ARIMA18

5.2.1 Mô hình dự đoán chuỗi thời gian Prophet19

PHẦN 6 – THỰC NGHIỆM, KẾT QUẢ, THẢO LUẬN20

6.1 Phân tích đa biến dùng K-Means Clustering (Phân cụm K-means)20

6.2 Mô hình dự đoán chuỗi thời gian21

6.2.1 Mô hình dự đoán chuỗi thời gian ARIMA21

6.2.2 Mô hình dự đoán chuỗi thời gian Prophet22

PHẦN 7 – KẾT LUẬN26
PHẦN 8 – PHỤ LỤC26
PHẦN 9 – ĐÓNG GÓP26
PHẦN 10 – THAM KHẢO27

PHẦN 1 – TÓM TẮT

Đề tài "Trực quan, phân tích và dự đoán dữ liệu bán hàng của một siêu thị tại thị trường Mỹ" nhằm mục đích nghiên cứu và phân tích dữ liệu bán hàng của một siêu thị tại Mỹ trong vòng 4 năm để đưa ra những kết luận hữu ích giúp cải thiện hiệu quả kinh doanh. Nghiên cứu sử dụng các phương pháp trực quan hóa dữ liệu, phân tích cụm, và dự đoán thống kê để khai thác thông tin từ dữ liệu bán hàng và đưa ra các dự đoán về xu hướng tiêu dùng và nhu cầu của khách hàng.

Trong đề tài này , các phương pháp phân tích và trực quan hóa dữ liệu được nhóm sử dụng gồm:

- Phân tích cụm: Sử dụng để phân tích khách hàng của siêu thị thành các nhóm tương đồng nhau dựa trên các thông tin như độ tuổi, giới tính, địa điểm và hành vi mua hàng. Sau đó, các nhóm khách hàng này có thể được so sánh với nhau để xem có sự khác biệt đáng kể về các chỉ số kinh doanh như doanh số, lợi nhuận, số lượng sản phẩm được bán ra, tỷ lệ chuyển đổi và mức độ trung thành của khách hàng.
- Trực quan hóa dữ liệu: Sử dụng các công cụ như biểu đồ, đồ thị để hiển thị dữ liệu bán hàng dưới dạng hình ảnh rõ ràng, giúp nhà quản lý có cái nhìn tổng quan về tình hình bán hàng và dễ dàng phát hiện ra các xu hướng và mô hình tiêu thụ của khách hàng.
- Dự đoán thống kê: Sử dụng các phương pháp dự đoán như hồi quy tuyến tính, mô hình dự đoán chuỗi thời gian để đưa ra dự đoán về doanh số bán hàng trong tương lai dựa trên các mẫu và xu hướng trong dữ liệu quá khứ.

Kết quả của bài nghiên cứu này sẽ giúp định hướng chiến lược kinh doanh cho siêu thị thời gian đó, đồng thời cung cấp thông tin hữu ích cho các nhà quản lý và nhà đầu tư trong lĩnh vực bán lẻ tại Mỹ.

PHẦN 2 – GIỚI THIỆU

Đề tài sẽ sử dụng các phương pháp phân tích và trực quan hóa dữ liệu để hiểu rõ hơn về sự phân bố, phân tích và tương quan giữa các yếu tố ảnh hưởng đến doanh số bán hàng của siêu thị, bao gồm: ngày, tháng, mùa, các ngày lễ, giá cả, phương thức vận chuyển, phân khúc khách hàng, loại sản phẩm, và các yếu tố kinh tế khác.

Ngoài ra, đề tài cũng sử dụng các thuật toán dự đoán chuỗi thời gian như ARIMA, Prophet để dự đoán doanh số bán hàng tương lai của siêu thị, giúp các nhà quản lý có thể đưa ra các quyết định kinh doanh phù hợp để tăng trưởng doanh số và cải thiện kinh doanh.

Trong báo cáo này, nhóm sử dụng Input của bài toán là tập các :

- Mã đơn hàng
- Ngày đặt hàng
- Ngày nhận hàng
- Phương thức vận chuyển
- Mã khách hàng
- Tên khách hàng
- Phân khúc khách hàng
- Quốc gia
- Thành phố
- Tiểu bang
- Mã bưu điện
- Vùng miền
- Mã sản phẩm
- Loại sản phẩm
- Loại sản phẩm phụ
- Tên sản phẩm
- Giá bán

Nhóm sử dụng những thuật toán:

- K-Means Clustering (Phân cụm K-means): Sử dụng để phân cụm các dữ liệu bán hàng thành các nhóm tương đồng với nhau, giúp nhìn rõ hơn các đặc tính của khách hàng và thị trường bán hàng.

- ARIMA (Auto-Regressive Integrated Moving Average): Sử dụng để mô hình hóa và dự đoán chuỗi thời gian, có thể giúp đưa ra dự đoán về doanh số bán hàng trong tương lai.
- Prophet: là một mô hình dự đoán chuỗi thời gian được phát triển bởi Facebook. Mô hình sử dụng một phương pháp phân tích chuỗi thời gian mới, gọi là mô hình tuyến tính theo ngày, để đưa ra dự đoán. Sử dụng Prophet ta có thể phân tích chuỗi thời gian về doanh số bán hàng của siêu thị, từ đó đưa ra các thông tin về xu hướng tăng giảm, định hướng và dự báo cho tương lai.

PHẦN 3 – DỮ LIỆU

Trong đề tài này, nhóm xin dùng bộ dữ liệu *Superstore Sales Dataset* để trực quan hoá mối liên hệ và tác động kể trên để phần nào hiểu rõ hơn về việc việc phân tích dữ liệu.

Theo như tác giả trên kaggle, tập dữ liệu xuất hiện trong nhóm người dùng tháng 12 của cộng đồng tableau. Hiện tại thì nhóm này đã không còn hoạt động và vẫn không tìm được tác giả của tập dữ liệu này là ai, nó xuất phát từ công ty hay doanh nghiệp nào và cũng không biết cách mà tập dữ liệu được thu thập.

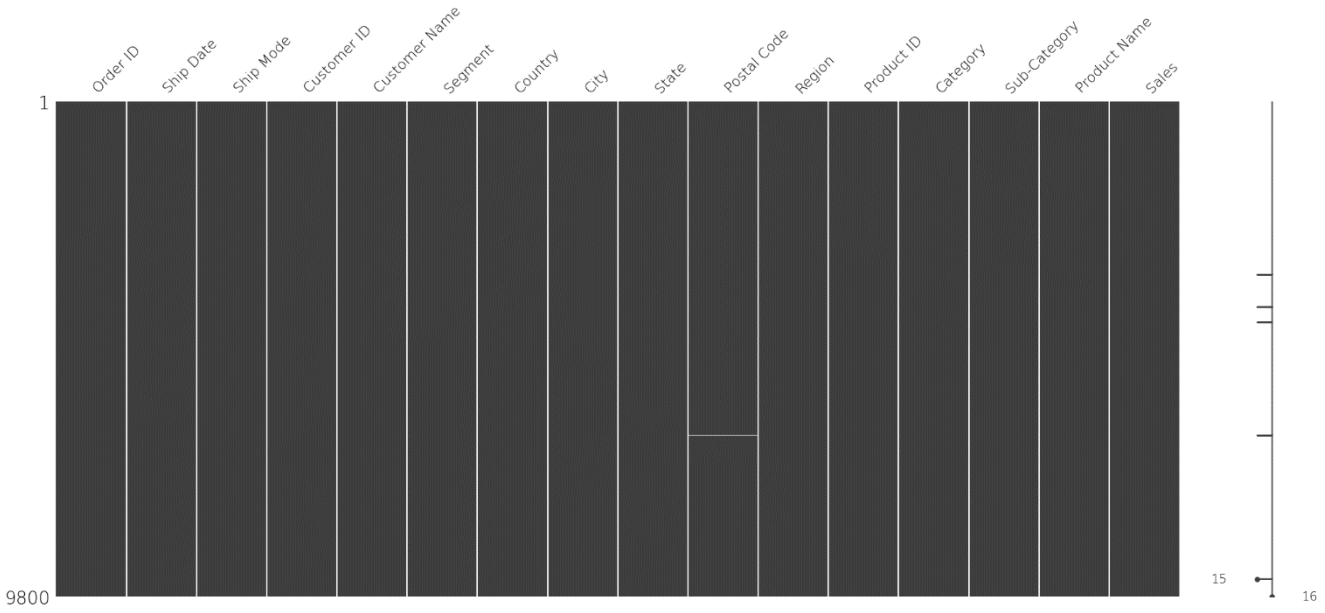
Bộ dữ liệu chứa hơn 9800 hàng, đã được tác giả kaggle thực hiện tiền xử lý, mỗi hàng đại diện cho một giao dịch bán hàng.

Bộ dữ liệu gồm 18 cột:

- Row ID: Số thứ tự
- Order ID: Mã đơn hàng
- Product ID: Mã định danh duy nhất cho mỗi sản phẩm.
- Order Date: Ngày đặt hàng.
- Ship Date: Ngày nhận hàng
- Ship Mode: Phương thức vận chuyển
- Sales: Giá bán cho mỗi sản phẩm.
- Customer ID: Mã định danh duy nhất cho mỗi khách hàng.
- Customer Name: Tên khách hàng
- Segment: Phân khúc của khách hàng, chẳng hạn như bán lẻ hoặc bán buôn.
- Country: Quốc gia nơi khách hàng đặt hàng.
- City: Thành phố nơi khách hàng đặt hàng.
- State: Tiểu bang hoặc tỉnh nơi khách hàng đặt hàng.
- Postal Code: Mã bưu chính nơi khách hàng đặt hàng.
- Region: Khu vực nơi khách hàng đặt hàng.
- Product ID: Mã sản phẩm
- Category: Loại sản phẩm, chẳng hạn như đồ nội thất hoặc công nghệ.
- Sub-Category: Danh mục con của sản phẩm, chẳng hạn như bàn hoặc laptop.
- Product name: Tên sản phẩm

3.1 Tiền xử lý

Qua một vài bước chuẩn hóa dữ liệu ngày tháng, loại bỏ những cột không cần thiết, thì ta thực hiện trực quan hóa dữ liệu bị thiếu như hình

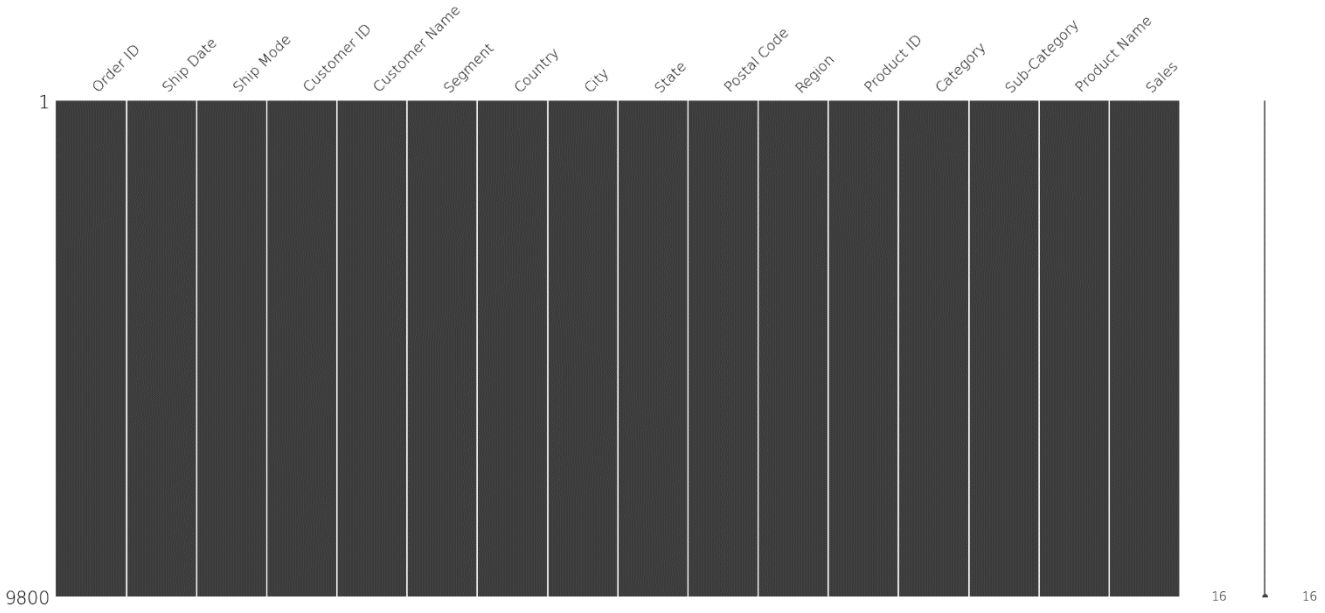


Ta có thể thấy ở dữ liệu một ít dữ liệu bị thiếu (missing value) ở những cột Postal Code.

Thì để xử lý những giá trị bị thiếu này, ta sẽ không bỏ nguyên bản ghi này mà mà thay vào đó ta sẽ thêm Postal code cho từng thành phố tương ứng

1. Đầu tiên thì ta cần tìm những thành phố có postal code bị thiếu
2. Thực hiện điền postal code tương ứng với thành phố ở từng hàng bị thiếu giá trị

Ta có thể thấy giá trị mã bưu chính bị thiếu ở đây chỉ có thành phố Burlington ở bang Vermont. Ta có mã bưu chính của thành phố này là 5401.

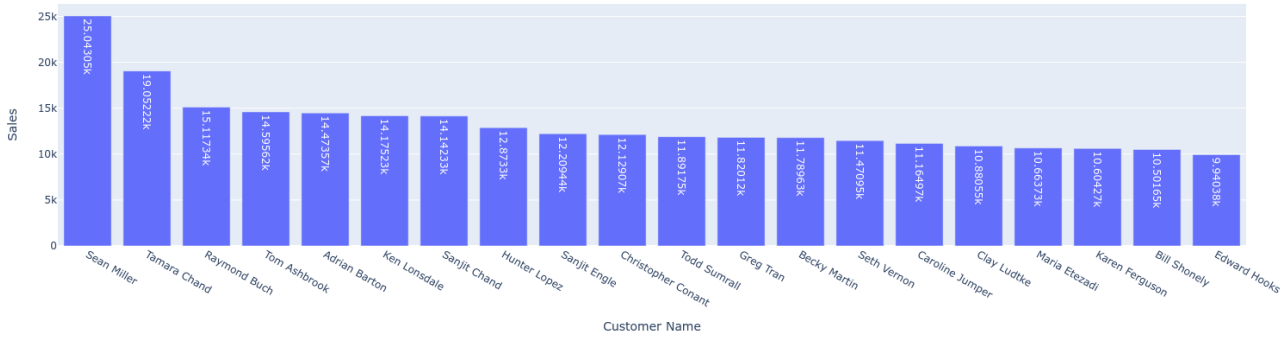


Có thể thấy dữ liệu sau khi được xử lí đã không còn dữ liệu trống và trở thành một bộ dữ liệu hoàn chỉnh.

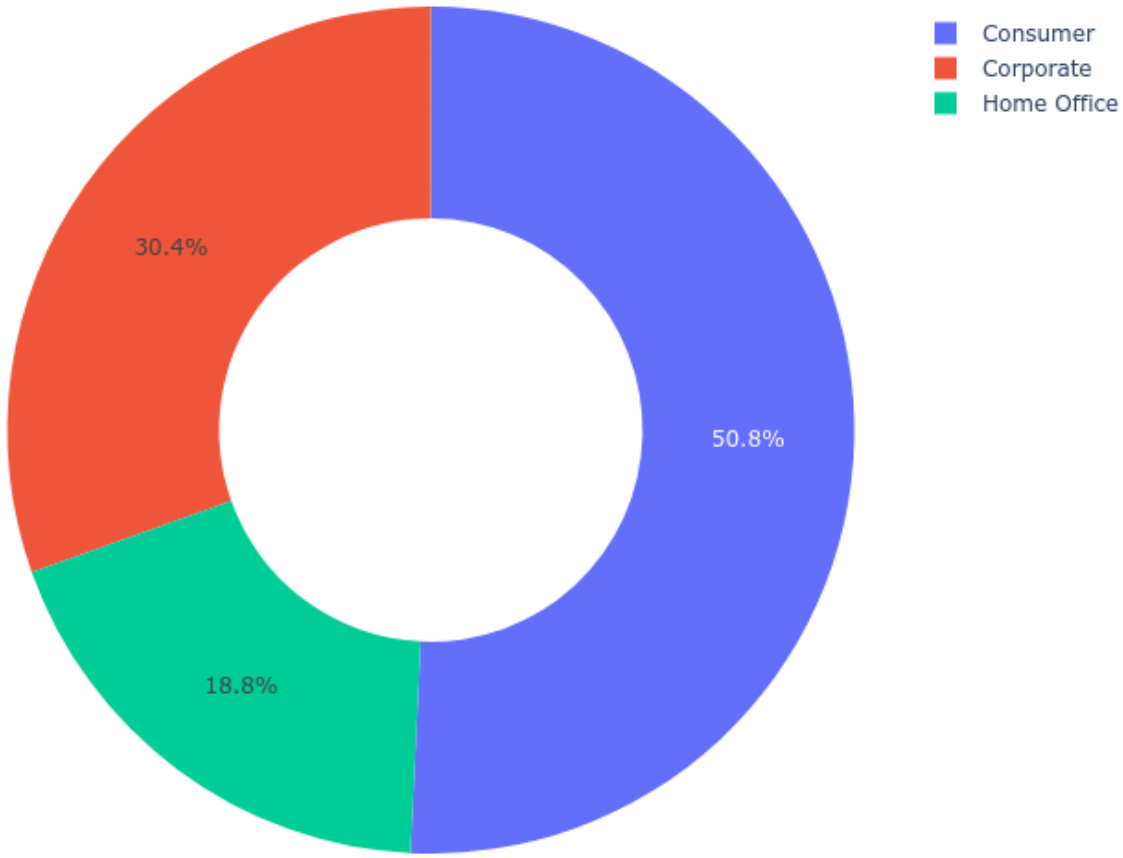
PHẦN 4 – TRỰC QUAN HOÁ DỮ LIỆU

4.1 Khách hàng:

4.1.1. Những khách hàng có doanh số cao nhất

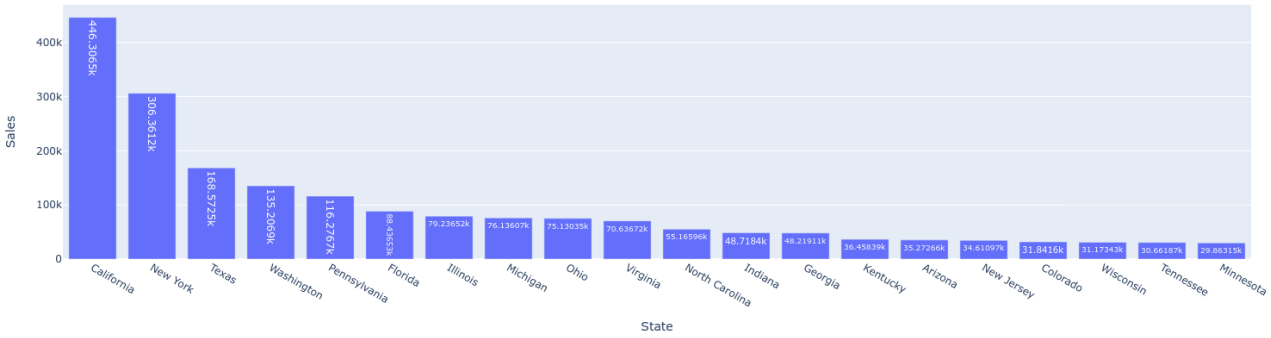


4.1.2 Tỷ lệ doanh thu theo từng phân khúc khách hàng

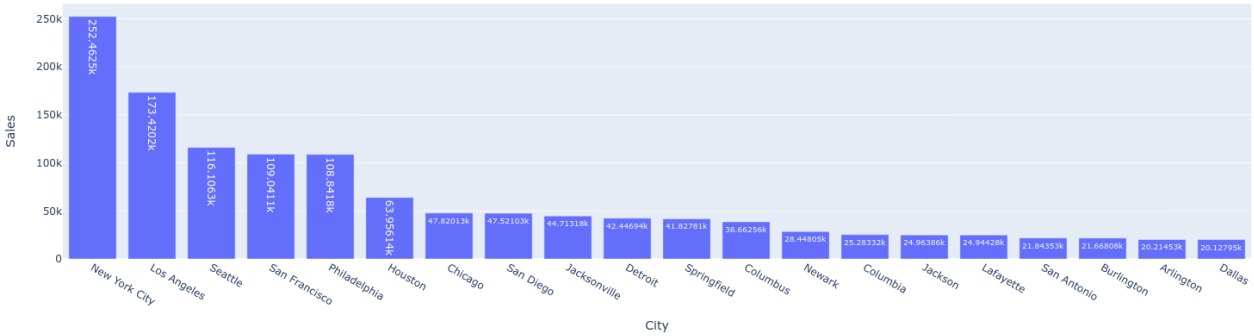


4.2 Địa chỉ khách hàng:

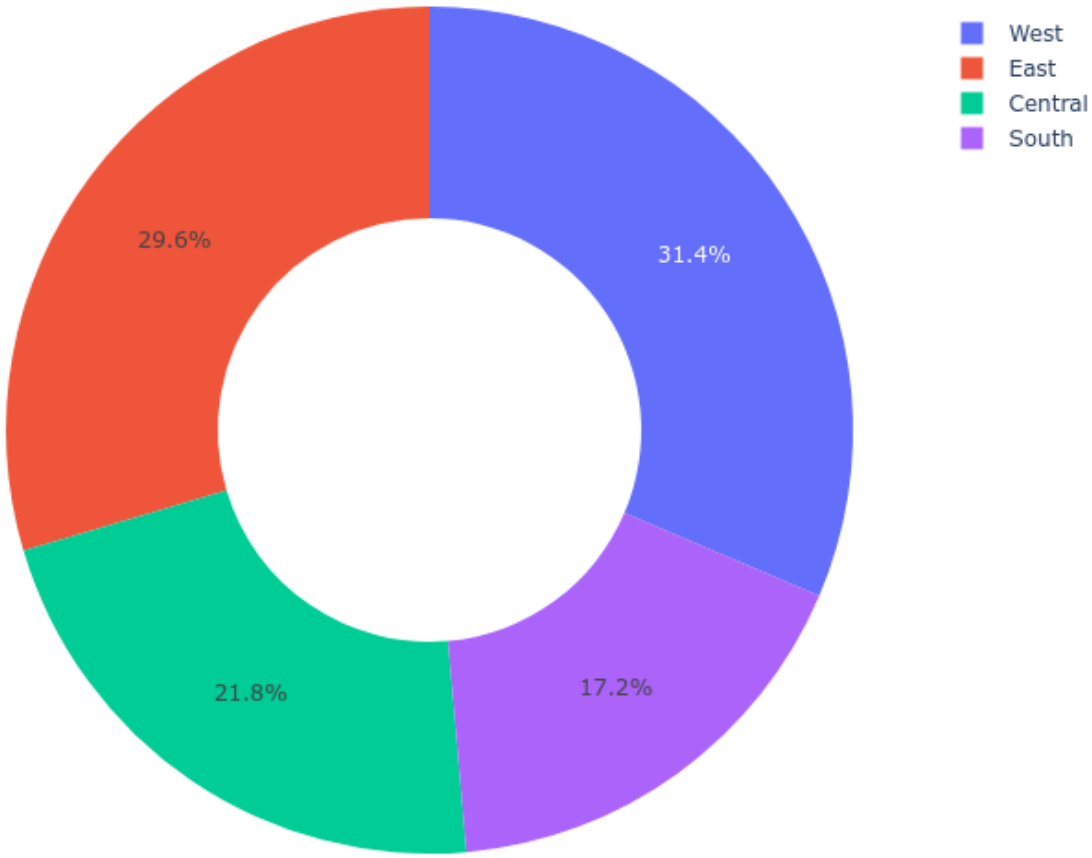
4.2.1. Những tiểu bang có doanh số cao nhất



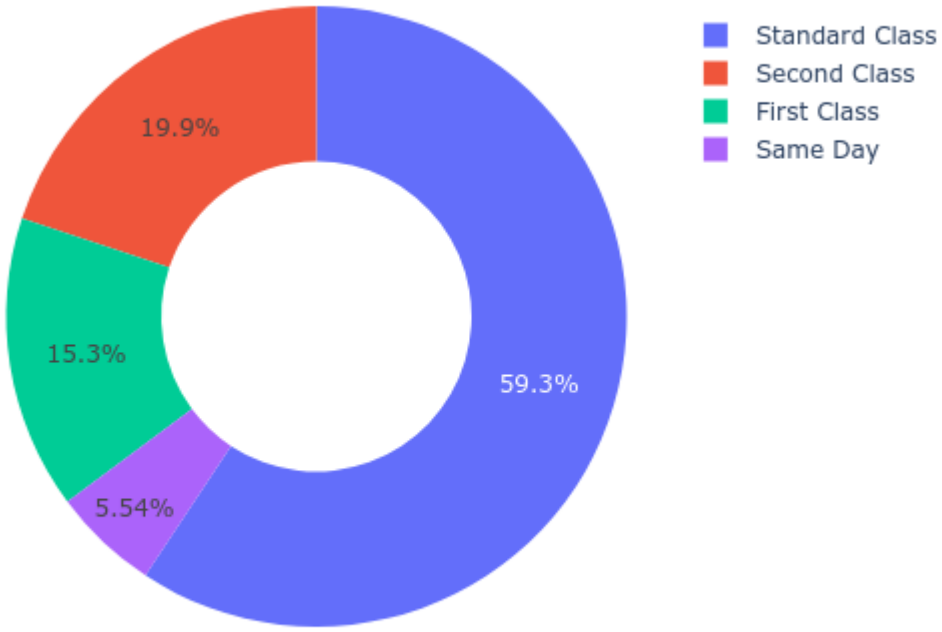
4.2.2. Những thành phố có doanh số cao nhất



4.2.3. Tỷ lệ doanh thu theo từng vùng miền

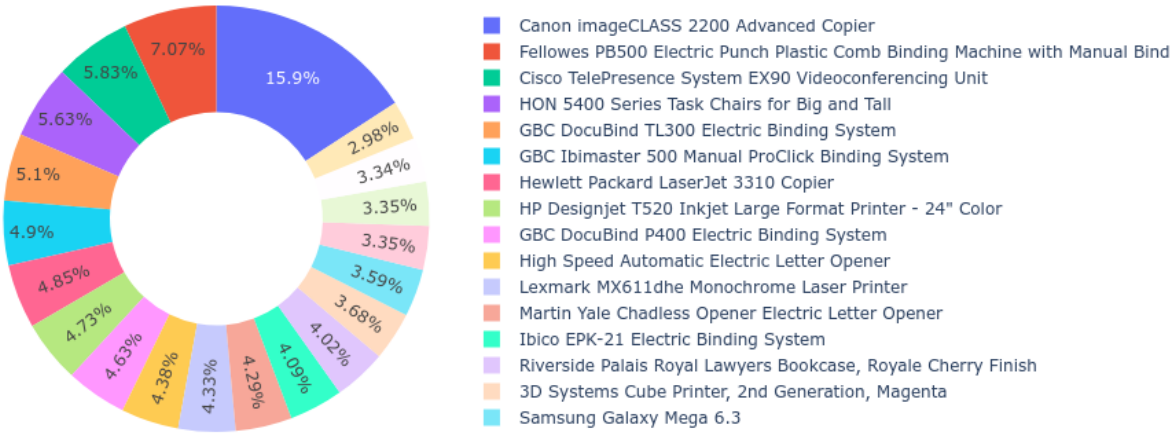


4.2.4. Tỷ lệ doanh thu theo từng phương thức vận chuyển

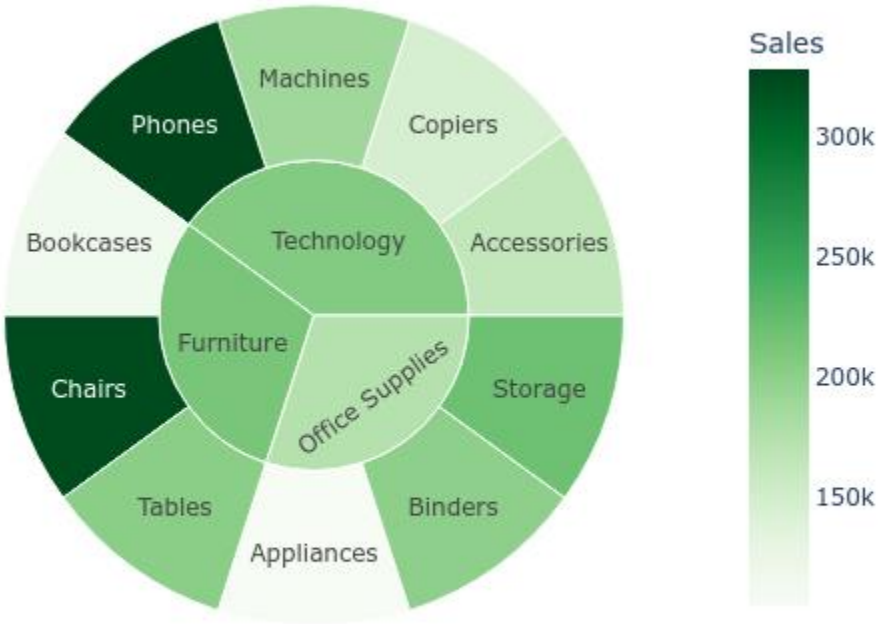


4.3 Sản phẩm:

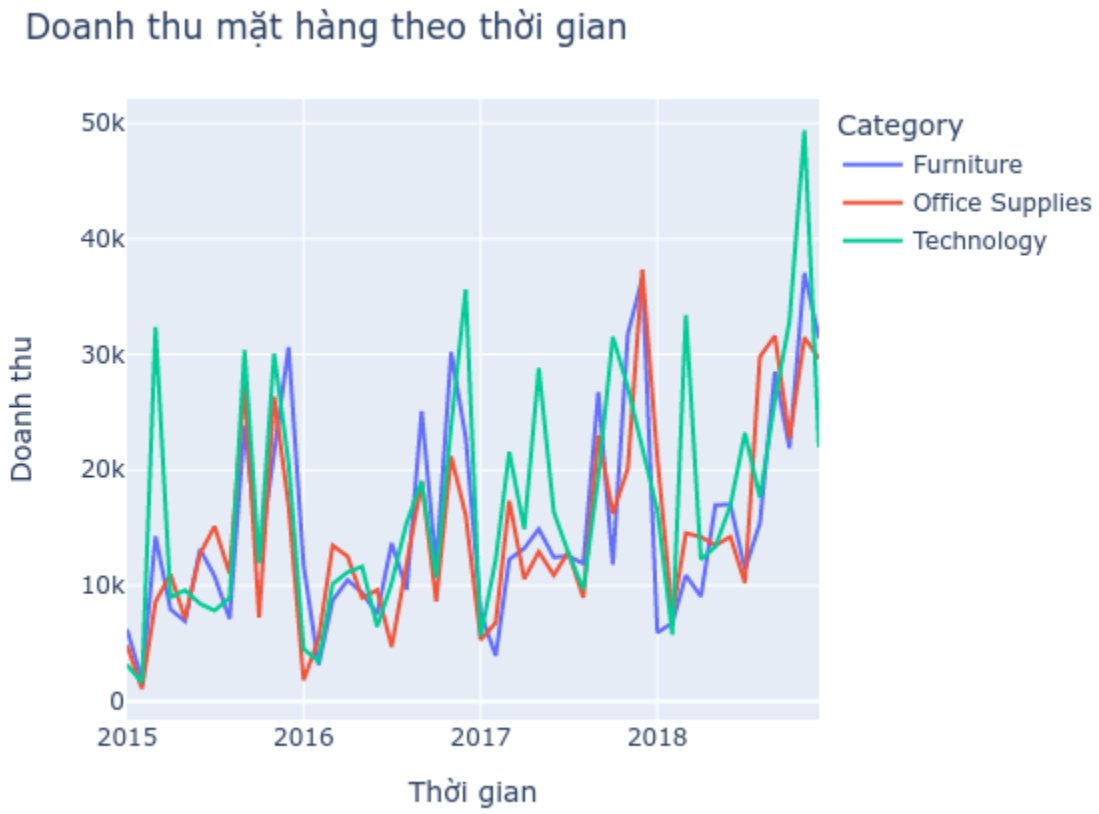
4.3.1. Những sản phẩm có doanh số cao nhất



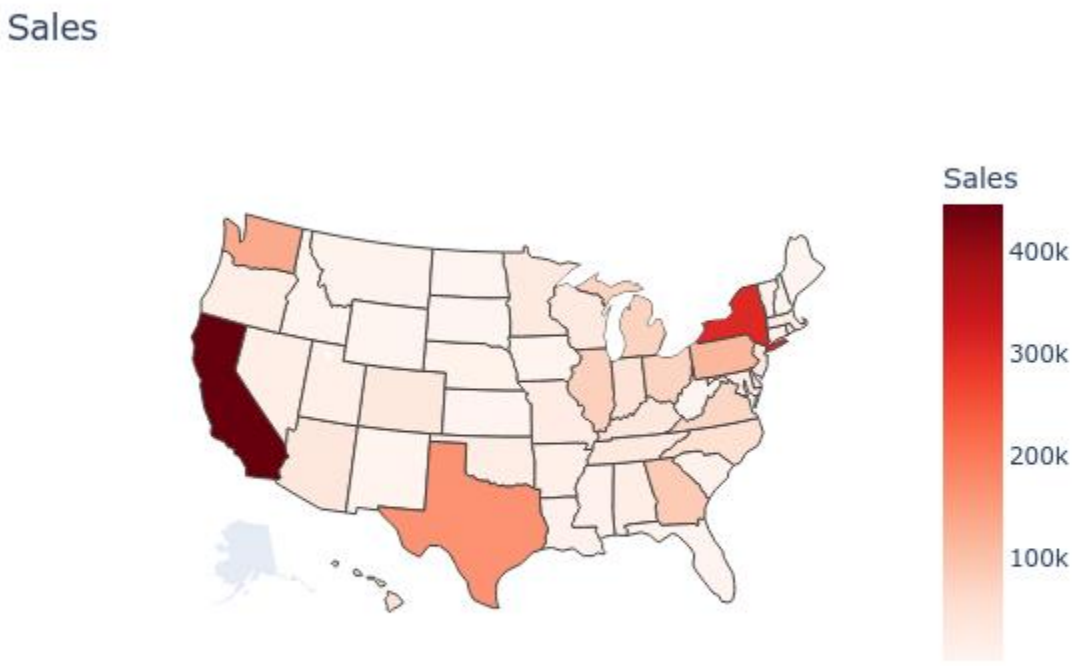
4.3.2. Tỷ lệ doanh thu theo từng danh mục sản phẩm, danh mục phụ sản phẩm



4.4 Doanh số bán được theo từng năm của từng loại danh mục sản phẩm:



4.5 Doanh số bán được của từng bang trên bản đồ nước Mỹ :


Sales
Sales
400k
300k
200k
100k

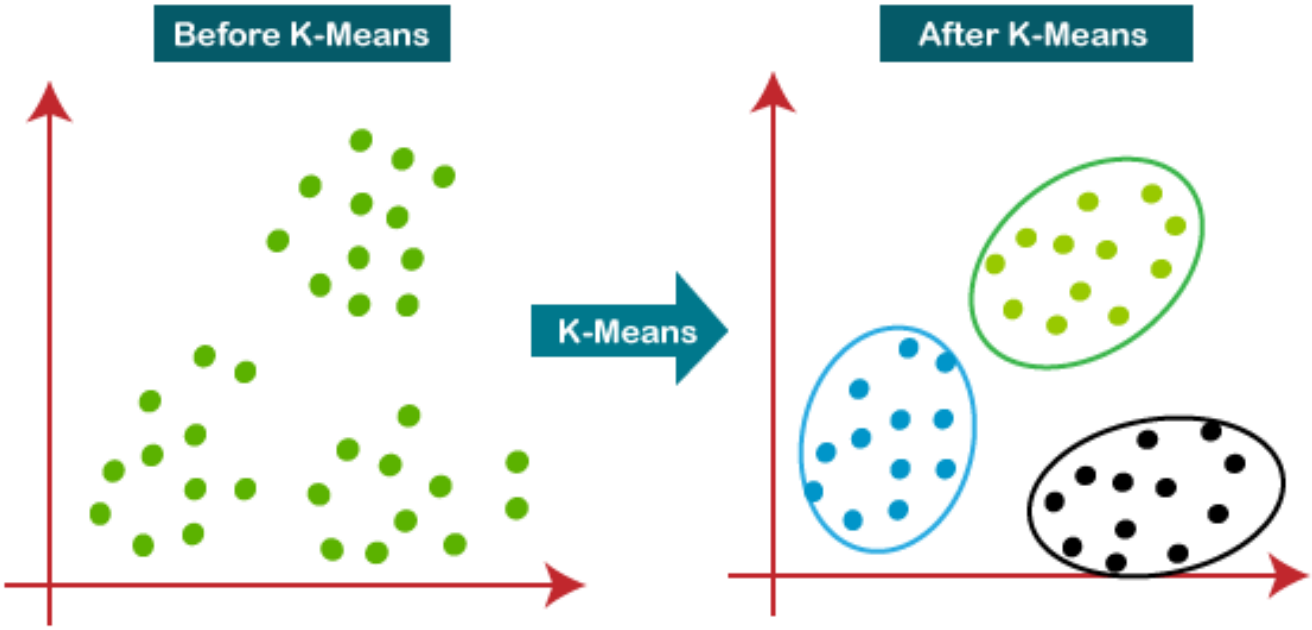
PHẦN 5 - MÔ HÌNH HÓA DỮ LIỆU

5.1 Phân tích cụm dùng K-Means Clustering (Phân cụm K-means)

K-means clustering là một phương pháp phân cụm dữ liệu phổ biến trong machine learning và thống kê. Phương pháp này nhằm phân chia tập dữ liệu thành các cụm (cluster) tương đồng nhau dựa trên độ tương đồng giữa các điểm dữ liệu. K-means clustering được áp dụng rộng rãi trong các bài toán khai thác dữ liệu, phân loại ảnh, phân tích chuỗi thời gian, và phân tích cụm khách hàng, v.v.

Quá trình phân cụm trong K-means diễn ra như sau:

1. Xác định số lượng cụm K cần phân chia.
2. Chọn ngẫu nhiên K điểm trong dữ liệu làm tâm cụm ban đầu.
3. Gán từng điểm dữ liệu vào cụm gần nhất với nó, dựa trên khoảng cách Euclidean giữa điểm và tâm cụm.
4. Cập nhật tâm cụm bằng cách lấy trung bình cộng của tất cả các điểm dữ liệu trong cụm.
5. Lặp lại các bước 3 và 4 cho đến khi các tâm cụm không thay đổi nữa hoặc đạt được số lần lặp tối đa.



Kết quả của K-means là các cụm tương đồng nhau dựa trên khoảng cách giữa các điểm dữ liệu và các tâm cụm. Tuy nhiên, việc lựa chọn số lượng cụm phù hợp là một vấn đề quan trọng và có thể ảnh hưởng đến kết quả cuối cùng.

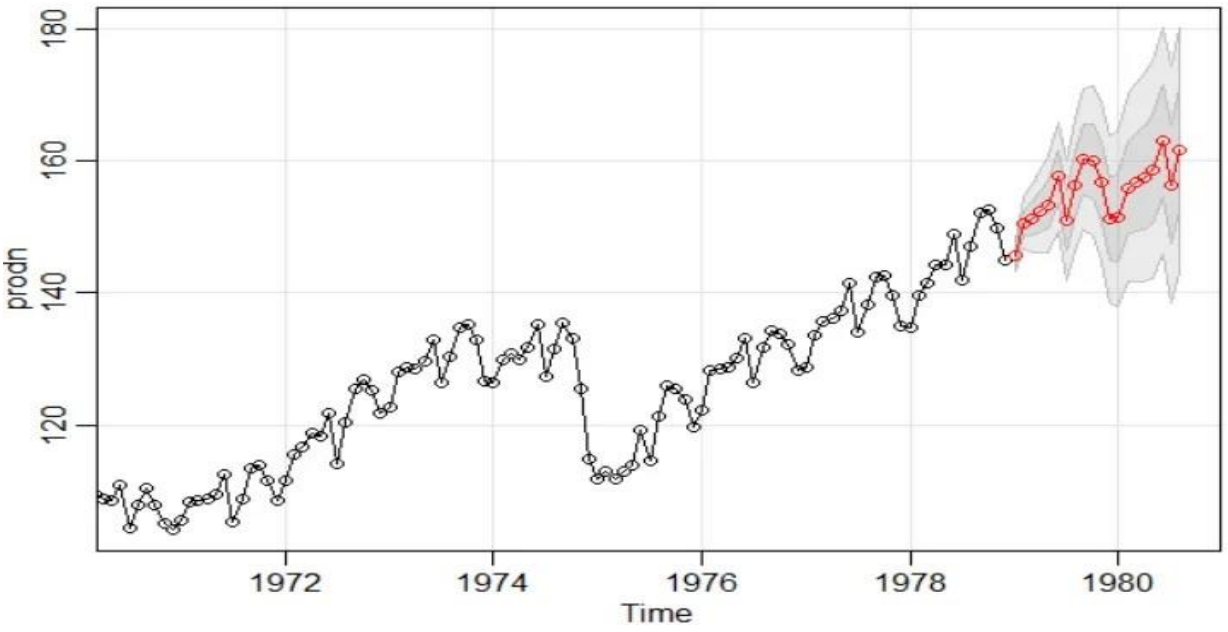
5.2 Mô hình dự đoán chuỗi thời gian

5.2.1 Mô hình dự đoán chuỗi thời gian SARIMA

SARIMA (Seasonal Autoregressive Integrated Moving Average) là một mô hình dự báo chuỗi thời gian phổ biến trong thống kê. SARIMA kết hợp các thành phần của mô hình ARIMA (Autoregressive Integrated Moving Average) với các thành phần mùa hóa, để xử lý các chu kỳ mùa vụ trong dữ liệu chuỗi thời gian.

Mô hình SARIMA có thể được biểu diễn bởi bốn tham số chính: p, d, q và P, D, Q, m. Trong đó:

- p: số lượng lags (khoảng thời gian trễ) của phần tự hồi quy (autoregressive) của mô hình. Tức là, số lượng giá trị quá khứ của chuỗi thời gian sẽ được sử dụng để dự báo giá trị tương lai.
- d: số lần khác biệt (difference) được áp dụng trên chuỗi thời gian để loại bỏ sự biến động của nó.
- q: số lượng lags của phần trung bình trượt (moving average) của mô hình. Tức là, số lượng giá trị trung bình của chuỗi thời gian sẽ được sử dụng để dự báo giá trị tương lai.
- P, D, Q: tương tự như p, d, q nhưng áp dụng cho các thành phần mùa hóa của chuỗi thời gian.
- m: số lượng chu kỳ trong một năm. Tức là, số lượng bước thời gian giữa các giá trị của chuỗi thời gian được xem như một chu kỳ mùa vụ.



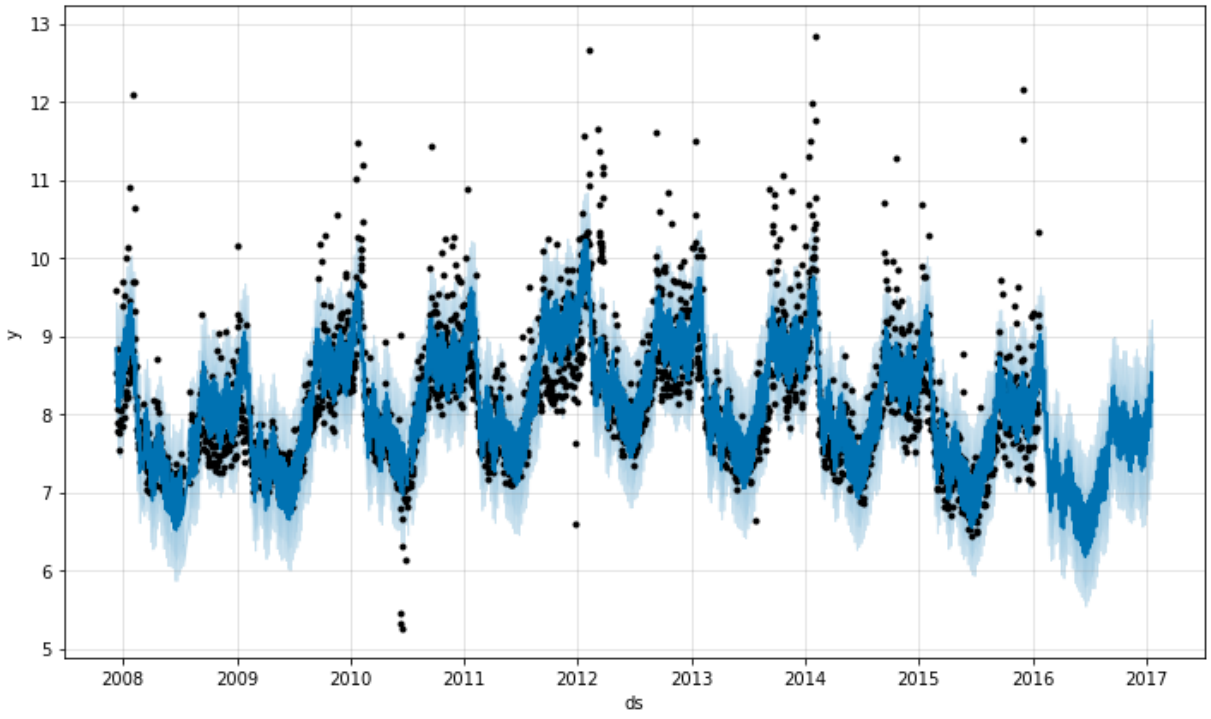
5.2.2 Mô hình dự đoán chuỗi thời gian Prophet

Prophet là một thư viện mã nguồn mở được phát triển bởi Facebook để dự báo chuỗi thời gian. Thư viện này có thể được sử dụng để dự báo các chuỗi thời gian có tính chất phức tạp, nhưng không đòi hỏi phải có nhiều kiến thức về thống kê.

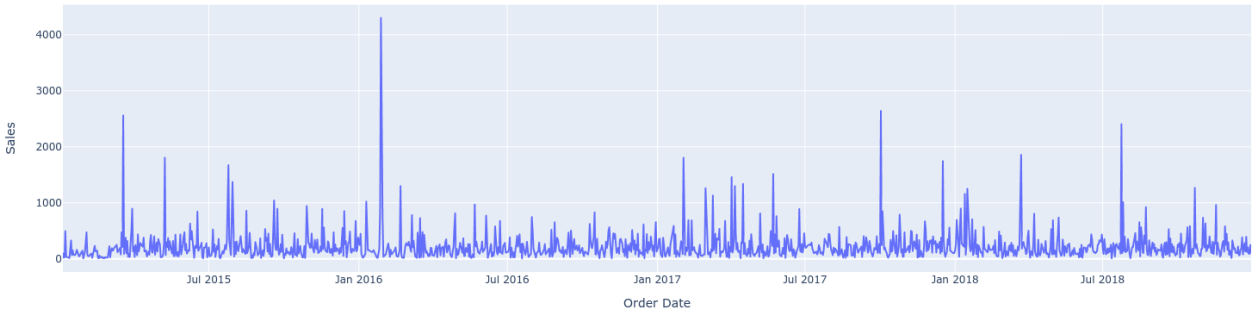
Prophet sử dụng một mô hình dự báo tuyến tính với một số tính năng đặc biệt, bao gồm:

- Tính toán các thành phần chính của chu kỳ mùa vụ trong dữ liệu chuỗi thời gian nhưng không yêu cầu người dùng chỉ định trước chu kỳ mùa vụ.
- Xử lý các giá trị bị khuyết (missing values) và giá trị ngoại lai (outliers) trong dữ liệu chuỗi thời gian.
- Tự động chuyển đổi các dữ liệu chuỗi thời gian không đồng nhất thành dữ liệu đồng nhất.

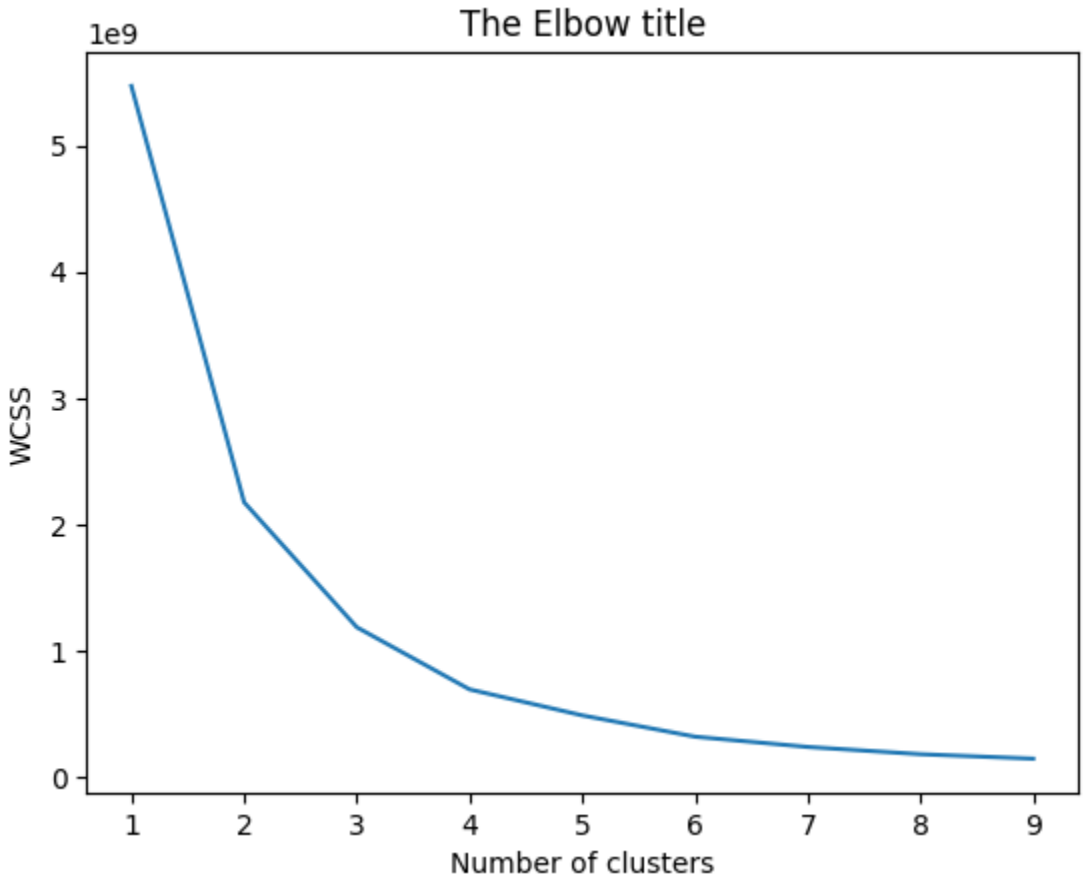
Prophet cũng cho phép người dùng tùy chỉnh các tham số mô hình nhưng vẫn giữ được tính đơn giản và trực quan của thư viện. Nó cũng cung cấp các công cụ trực quan hóa để hiển thị dữ liệu chuỗi thời gian và các dự đoán.

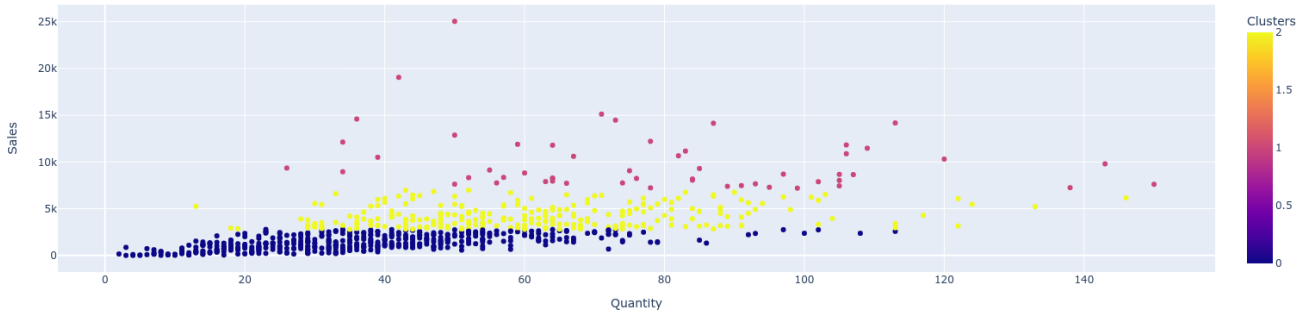


PHẦN 6 – THỰC NGHIỆM, KẾT QUẢ, THẢO LUẬN



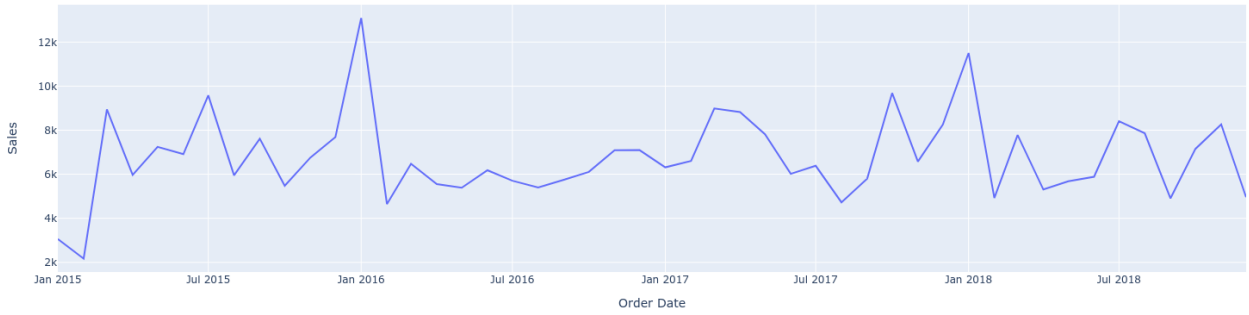
6.1 Phân tích cụm dùng K-Means Clustering (Phân cụm K-means)

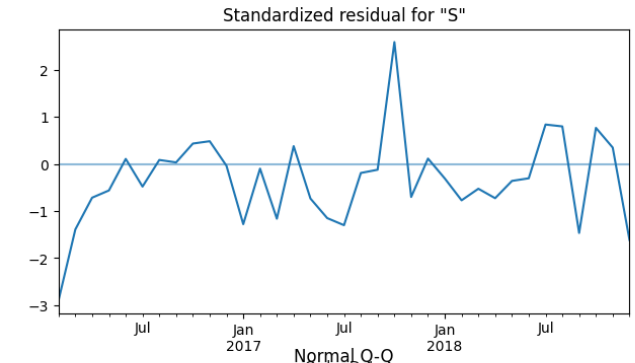


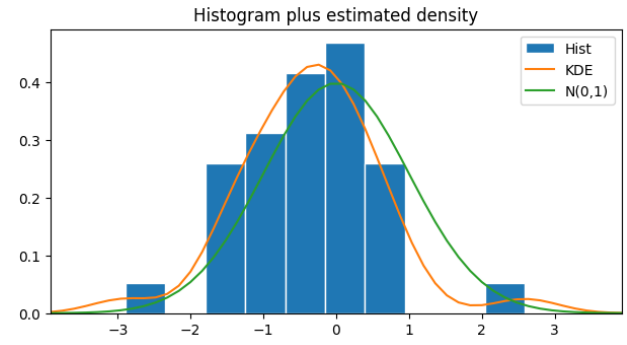


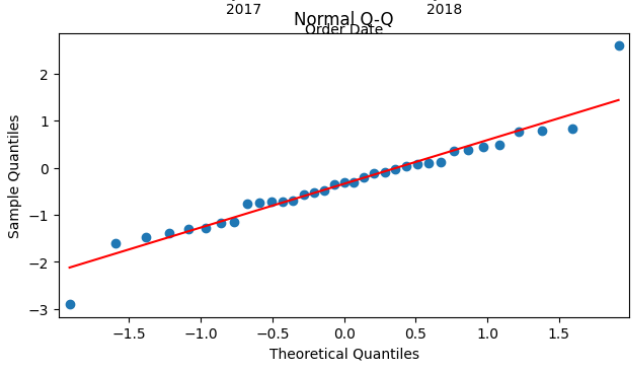
6.2 Mô hình dự đoán chuỗi thời gian

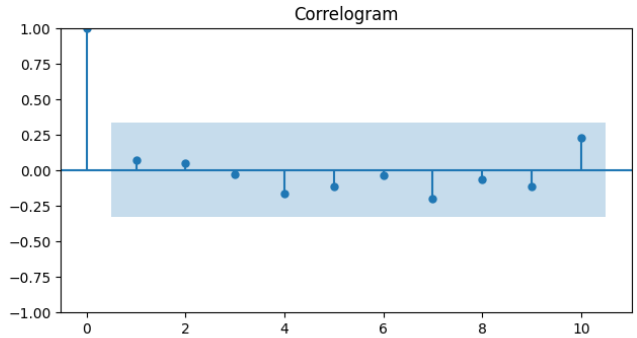
6.2.1 Mô hình dự đoán chuỗi thời gian ARIMA

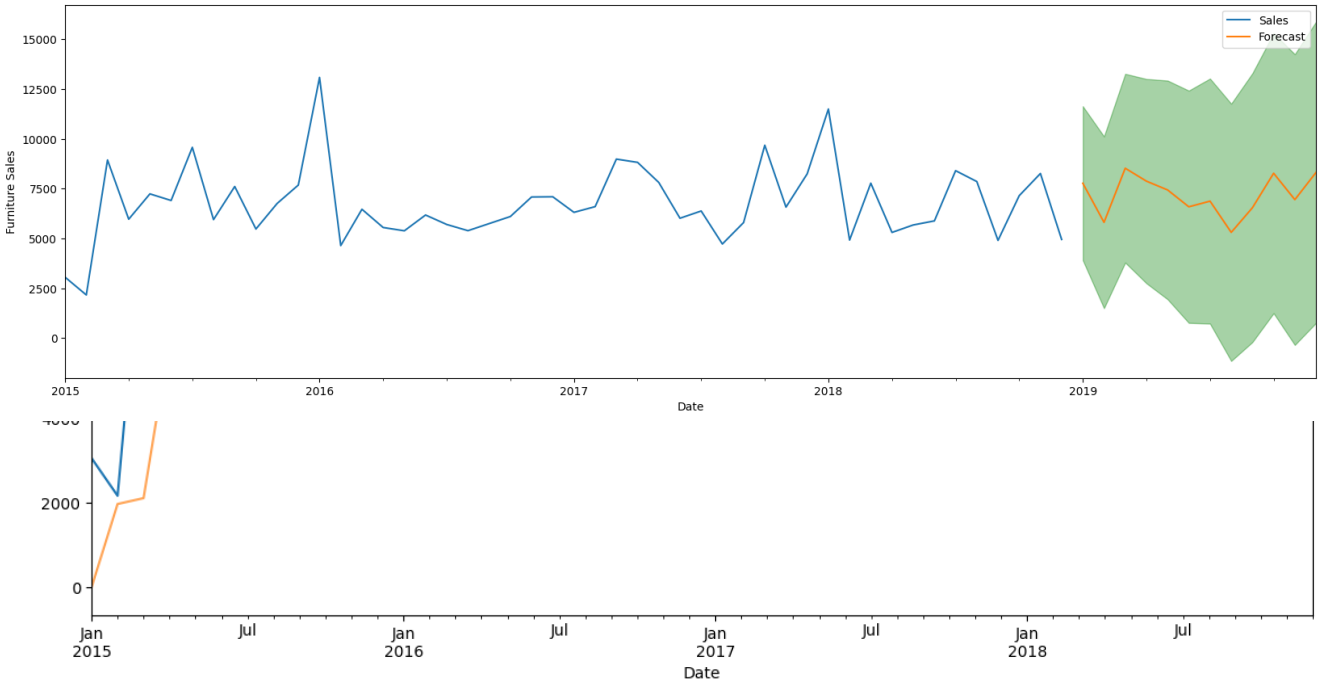




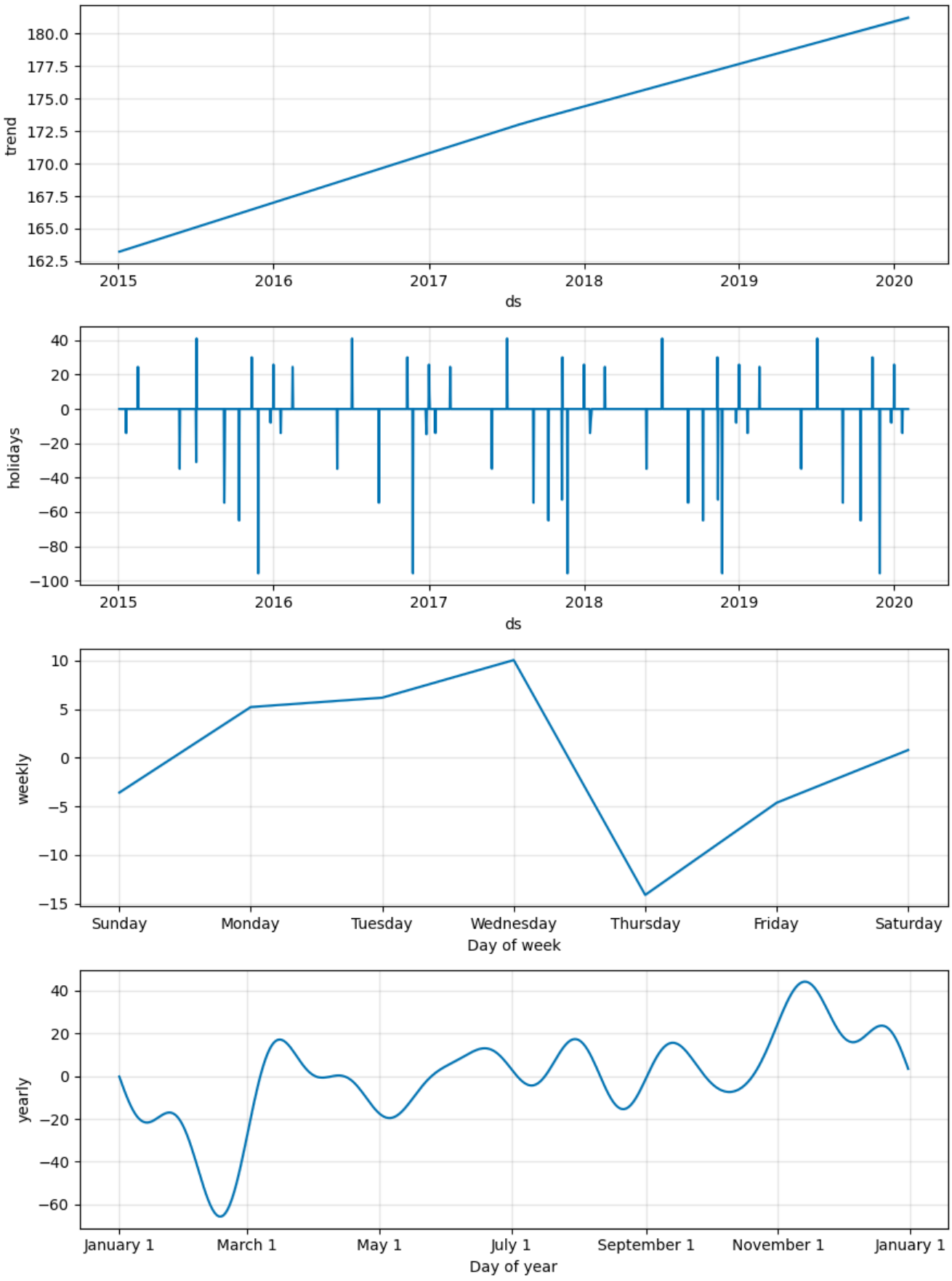


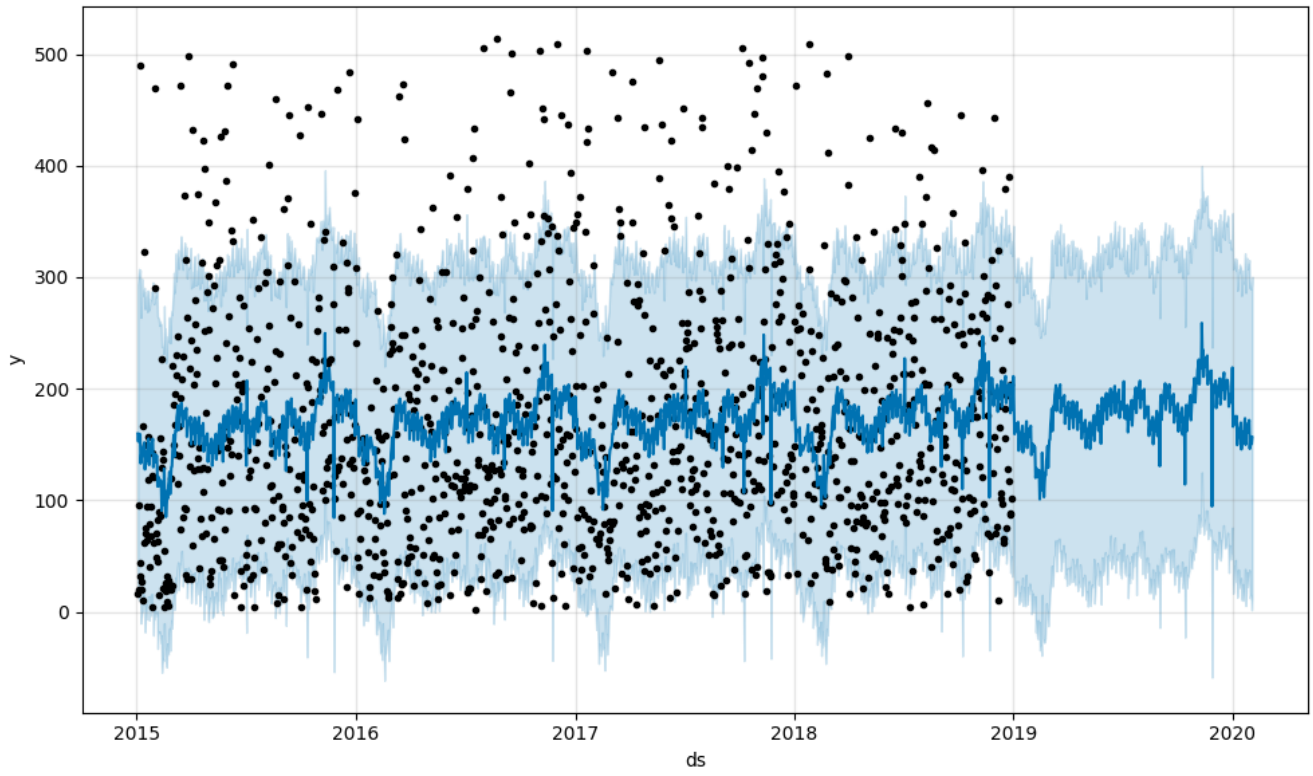






6.2.2 Mô hình dự đoán chuỗi thời gian Prophet


trend
180.0
177.5
175.0
172.5
170.0
167.5
165.0
162.5
2015
2016
2017
2018
2019
2020
ds
holidays
40
20
0
−20
−40
−60
−80
−100
2015
2016
2017
2018
2019
2020
ds
weekly
10
5
0
−5
−10
−15
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week
yearly
40
20
0
−20
−40
−60
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year


500
400
300
200
100
0
y
2015
2016
2017
2018
2019
2020
ds

PHẦN 7 – KẾT LUẬN

PHẦN 8 – PHỤ LỤC

PHẦN 9 – ĐÓNG GÓP

MSSV	Họ và Tên	Công việc	Hoàn thành
20133076	Văn Mai Thanh Nhật		100%
20133082	Huỳnh Minh Phước		100%
20133013	Trần Nguyên Hạnh		100%
20133035	Trần Đông		100%

PHẦN 10 – THAM KHẢO

1. Rohit Sahoo, Superstore Sales Dataset, Kaggle , 2023, đường dẫn: https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting.
2. Samruddhi Mhatre, Part 1: Exploratory Data Analysis, Kaggle , 2023, đường dẫn: https://www.kaggle.com/code/samruddhim/part-1-exploratory-data-analysis.
3. Samruddhi Mhatre, Part 2: Time Series Analysis, Kaggle , 2023, đường dẫn: https://www.kaggle.com/code/samruddhim/part-2-time-series-analysis.
4. OH SEOK KIM, Predicting Future by LSTM, Prophet, Neural Prophet, Kaggle, 2023, đường dẫn: https://www.kaggle.com/code/ohseokkim/predicting-future-by-lstm-prophet-neural-prophet.
5. Thư viện python: pandas, numpy, plotly, seaborn, statsmodel, matplotlib, missingno, sklearn