BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC UEH

ĐỒ ÁN CUỐI KỲ

BIỂU DIỄN TRỰC QUAN DỮ LIỆU

PHÂN TÍCH TÌNH HÌNH KINH DOANH CỦA MỘT SIÊU THỊ TẠI MỸ

Giảng viên bộ môn	**:**	Nguyễn An Tế
Lớp học phần	**:**	23C1INF50908201
Buổi học	**:**	Chiều T7
Nhóm	**:**	6

TP. HỒ CHÍ MINH, NĂM 2023

Mục lục

1. Tổng quan

1.1. Mục tiêu

Với nhu cầu mua sắm ngày càng tăng cao và sự cạnh tranh gay gắt trên thị trường buôn bán hàng hóa, việc khai thác các ý tưởng thông qua dữ liệu là một điều vô cùng cần thiết để nâng cao năng lực cạnh tranh đồng thời giúp các nhà quản trị nhắm đến mục tiêu tối đa hóa lợi nhuận cho doanh nghiệp.

Do vậy với dự án lần này nhóm em đặt mục tiêu nhằm tìm hiểu về tình hình kinh doanh của một cửa hàng, bằng cách quan sát thông qua từng nhóm biểu đồ khác nhau để có thể có được những thông tin hữu ích từ trong bộ dữ liệu này.

1.2. Mô tả bộ dữ liệu

"Sample-Superstore" là một bộ dữ liệu chứa các giao dịch của một của hàng bán lẻ ghi nhận từ đầu năm 2014 đến đầu năm 2018. Bộ dữ liệu có tổng cộng 9.994 bản ghi và 21 thuộc tính.

Nguồn: Sample - Superstore

1.3. Ý nghĩa thuộc tính

STT	Tên thuộc tính	Giải thích	Giá trị
1	Row ID	Chỉ số dòng	1 - 9994
2	Order ID	Mã đơn hàng	5009 giá trị
3	Order Date	Ngày đặt hàng	03/01/2014 - 30/12/2017
4	Ship Date	Ngày giao hàng	07/01/2014 - 05/01/2018
5	Ship Mode	Phương thức vận chuyển	- Standard Class - Second Class - First Class - Same day
6	Customer ID	Mã khách hàng	793 giá trị

7	Customer Name	Tên khách hàng	793 giá trị
8	Segment	Phân khúc khách hàng	- Consumer - Corporate - Home Office
9	Country	Quốc gia	United States
10	City	Thành phố	531 giá trị
11	State	Bang	49 giá trị
12	Postal Code	Mã bưu điện	
13	Region	Khu vực	- West - East - Central - South
14	Product ID	Mã sản phẩm	1862 giá trị
15	Category	Danh mục sản phẩm	- Office Supplies - Furniture - Technology
16	Sub-Category	Loại hình sản phẩm	17 giá trị
17	Product Name	Tên sản phẩm	1850 giá trị

18	Sales	Doanh thu	
19	Quantity	Số lượng	
20	Discount	Tỉ lệ chiết khẩu	0 - 0.8
21	Profit	Lợi nhuận	

2. Tiền xử lý

2.1. Môi trường triển khai ứng dụng

+ Tải các thư viện cần thiết:

```
!pip install pycountry
!pip install -U ridgeplot
```

+ Cài đặt các thư viện:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import plotly as pl
import pycountry
import warnings
import os as os
import scipy.stats as ss
from ridgeplot import ridgeplot
from datetime import datetime, timedelta
from itertools import product
from scipy.stats import pearsonr
```

+ Bỏ qua các cảnh báo:

```
np.seterr(divide = 'ignore')
warnings.filterwarnings('ignore')
```

+ Đọc bộ dữ liệu Sample – Superstore:

```
df = pd.read_csv('Sample - Superstore.csv', encoding='windows-1252')
```

2.2. Thăm dò dữ liệu

+ Kiểm tra thông tin bộ dữ liệu:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Row ID         9994 non-null   int64
 1   Order ID       9994 non-null   object
 2   Order Date     9994 non-null   object
 3   Ship Date      9994 non-null   object
 4   Ship Mode      9994 non-null   object
 5   Customer ID    9994 non-null   object
 6   Customer Name  9994 non-null   object
 7   Segment        9994 non-null   object
 8   Country        9994 non-null   object
 9   City           9994 non-null   object
 10  State          9994 non-null   object
 11  Postal Code    9994 non-null   int64
 12  Region         9994 non-null   object
 13  Product ID     9994 non-null   object
 14  Category       9994 non-null   object
 15  Sub-Category   9994 non-null   object
 16  Product Name   9994 non-null   object
 17  Sales          9994 non-null   float64
 18  Quantity       9994 non-null   int64
 19  Discount       9994 non-null   float64
 20  Profit         9994 non-null   float64
dtypes: float64(3), int64(3), object(15)
memory usage: 1.6+ MB
```

+ Trích mẫu bộ dữ liệu:

	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	...	Sales	Quantity	Discount	Profit	Order Year	Sales_log	Profit_log	Month	Log_Sales	Year
0	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	...	261.9600	2	0.00	41.9136	2016	5.568192	3.735610	11	5.572002	2016
1	CA-2016-152156	2016-11-08	2016-11-11	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	...	731.9400	3	0.00	219.5820	2016	6.595699	5.391726	11	6.597064	2016
2	CA-2016-138688	2016-06-12	2016-06-16	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	...	14.6200	2	0.00	6.8714	2016	2.682390	1.927368	6	2.748552	2016
3	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	...	957.5775	5	0.45	-383.0310	2015	6.864407	NaN	10	6.865450	2015
4	US-2015-108966	2015-10-11	2015-10-18	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	...	22.3680	2	0.20	2.5164	2015	3.107631	0.922829	10	3.151368	2015
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	CA-2014-110422	2014-01-21	2014-01-23	Second Class	TB-21400	Tom Boeckenhauer	Consumer	United States	Miami	Florida	...	25.2480	3	0.20	4.1028	2014	3.228747	1.411670	1	3.267590	2014
9990	CA-2017-121258	2017-02-26	2017-03-03	Standard Class	DB-13060	Dave Brooks	Consumer	United States	Costa Mesa	California	...	91.9600	2	0.00	15.6332	2017	4.521354	2.749397	2	4.532169	2017
9991	CA-2017-121258	2017-02-26	2017-03-03	Standard Class	DB-13060	Dave Brooks	Consumer	United States	Costa Mesa	California	...	258.5760	2	0.20	19.3932	2017	5.555190	2.964922	2	5.559050	2017
9992	CA-2017-121258	2017-02-26	2017-03-03	Standard Class	DB-13060	Dave Brooks	Consumer	United States	Costa Mesa	California	...	29.6000	4	0.00	13.3200	2017	3.387774	2.589267	2	3.421000	2017
9993	CA-2017-119914	2017-05-04	2017-05-09	Second Class	CC-12220	Chris Cortes	Consumer	United States	Westminster	California	...	243.1600	2	0.00	72.9480	2017	5.493720	4.289747	5	5.497824	2017

9994 rows × 26 columns

+ Để thuận tiện cho việc biểu diễn phía sau ta sẽ chuyển kiểu của “Order Date” và “Ship Date” sang datetime

```
## Chuyển đổi các cột kiểu dữ liệu thời gian sang datetime
df['Order Date'] = pd.to_datetime(df['Order Date'], format="%m/%d/%Y")
df['Ship Date'] = pd.to_datetime(df['Ship Date'], format="%m/%d/%Y")
```

+ Do bản chất postal code là mã bưu điện do đó nó không phải một biến định lượng do vậy ta cần chuyển nó về kiểu dữ liệu object để tiện cho phân tích phía sau:

```
## Chuyển đổi các cột sang dạng object
df['Postal Code'] = df['Postal Code'].astype('object')
```

+ Vì Row ID là cột chỉ chỉ số dòng trong bộ dữ liệu, nên ở đây ta sẽ loại bỏ cột Row ID khỏi bộ dữ liệu:

```
## Xóa cột 'Row ID'
df = df.drop(['Row ID'], axis=1)
```

+ Thống kê mô tả các biến định lượng:

	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000
mean	229.858001	3.789574	0.156203	28.656896
std	623.245101	2.225110	0.206452	234.260108
min	0.444000	1.000000	0.000000	-6599.978000
25%	17.280000	2.000000	0.000000	1.728750
50%	54.490000	3.000000	0.200000	8.666500
75%	209.940000	5.000000	0.200000	29.364000
max	22638.480000	14.000000	0.800000	8399.976000

+ Thống kê mô tả các biến phân loại:

	Order ID	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name
count	9994	9994	9994	9994	9994	9994	9994	9994	9994	9994	9994	9994	9994	9994
unique	5009	4	793	793	3	1	531	49	631	4	1862	3	17	1850
top	CA-2017-100111	Standard Class	WB-21850	William Brown	Consumer	United States	New York City	California	10035	West	OFF-PA-10001970	Office Supplies	Binders	Staple envelope
freq	14	5968	37	37	5191	9994	915	2001	263	3203	19	6026	1523	48

2.3. Làm sạch dữ liệu

2.3.1. Missing values

+ Kiểm tra các giá trị rỗng bằng hàm df.isnull() để tìm số giá trị rỗng trong các cột:

```
Order ID         0
Order Date       0
Ship Date        0
Ship Mode        0
Customer ID      0
Customer Name    0
Segment          0
Country          0
City             0
State            0
Postal Code      0
Region           0
Product ID       0
Category         0
Sub-Category     0
Product Name     0
Sales            0
Quantity         0
Discount         0
Profit           0
dtype: int64
```

Nhận xét: Bộ dữ liệu không có giá trị rỗng nào nên ta không cần thực hiện các bước xử lý rỗng cho bộ dữ liệu.

2.3.2. Noisy data

Để kiểm tra các giá trị ngoại lai, ta có thể sử dụng quy tắc 3 sigma để in ra số giá trị mà được xem là outlier trong mỗi biến như sau:

```
Cột Sales có 127 giá trị nhiễu là các giá trị ngoài khoảng (-1640, 2100)
Cột Quantity có 113 giá trị nhiễu là các giá trị ngoài khoảng (-3, 10)
Cột Discount: Không có outlier.
Cột Profit có 107 giá trị nhiễu là các giá trị ngoài khoảng (-674, 731)
```

Nhận xét: Sau khi nhận được kết quả in ra số giá trị ngoại lai (outlier) khi sử dụng quy tắc 3-sigma, ta nhận thấy, 3 thuộc tính “Sales”, “Profit” và “Quantity” đều xuất hiện outlier.

Tuy nhiên xét về ngữ cảnh của bài, ta thấy được có những mặt hàng có giá trị rất cao trong bộ dữ liệu như máy in, máy photocopy điều đó dẫn đến những hóa đơn có những mặt hàng này sẽ có thu được doanh thu và lợi nhuận cao hơn so với phần còn lại và theo quy tắc 3-sigma nó sẽ được nhận định là outlier.

Ngoài ra, cũng có những sản phẩm có giá thành thấp, do vậy việc mua với số lượng lớn ở những sản phẩm này cũng là điều dễ hiểu đặc biệt với nhóm khách hàng là các tổ chức.

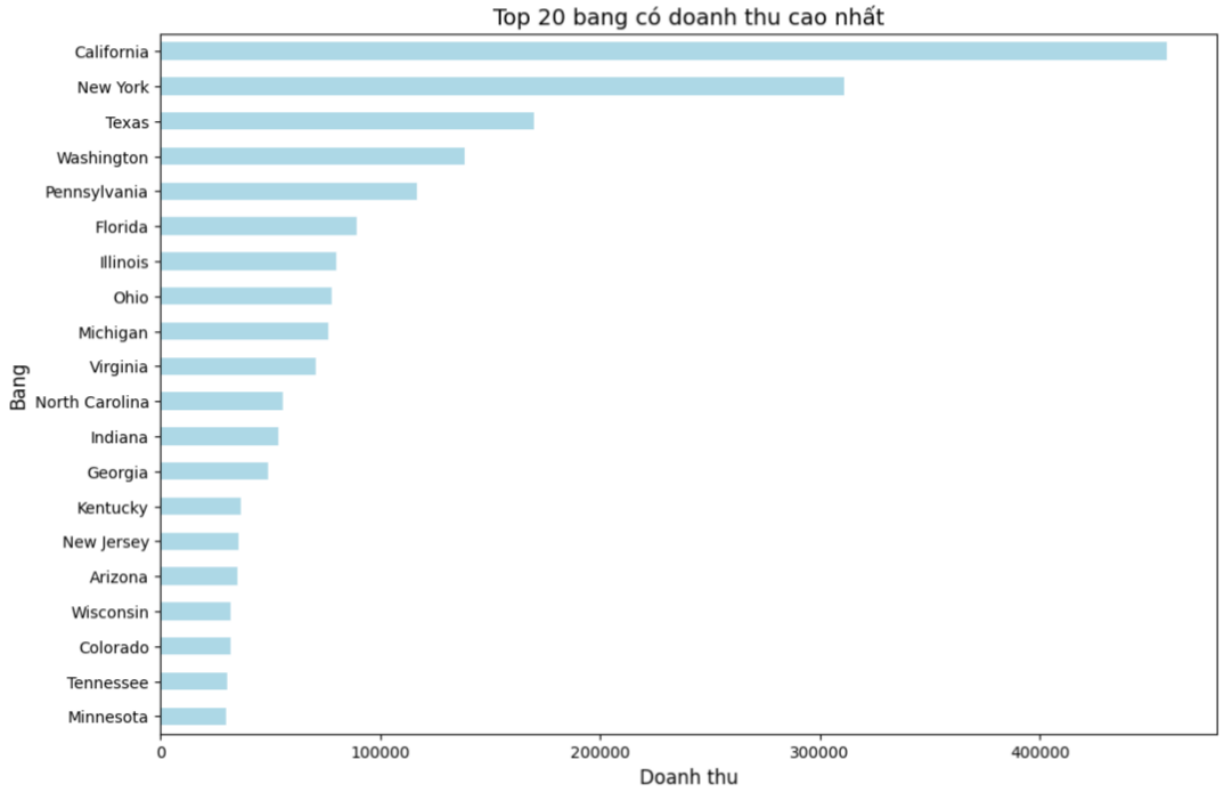
Đồng thời, với mục tiêu tìm hiểu tình hình kinh doanh thông qua biểu diễn trực quan bộ dữ liệu, việc xử lý nhiễu sẽ không quá cấp thiết, đồng thời giúp ta tránh việc mất hoặc sai lệch thông tin sự kiện mà cửa hàng đã thật sự trải qua.

=> Do vậy nhóm sẽ không xử lý các giá trị nhiễu trong đồ án này.

3. Biểu diễn trực quan

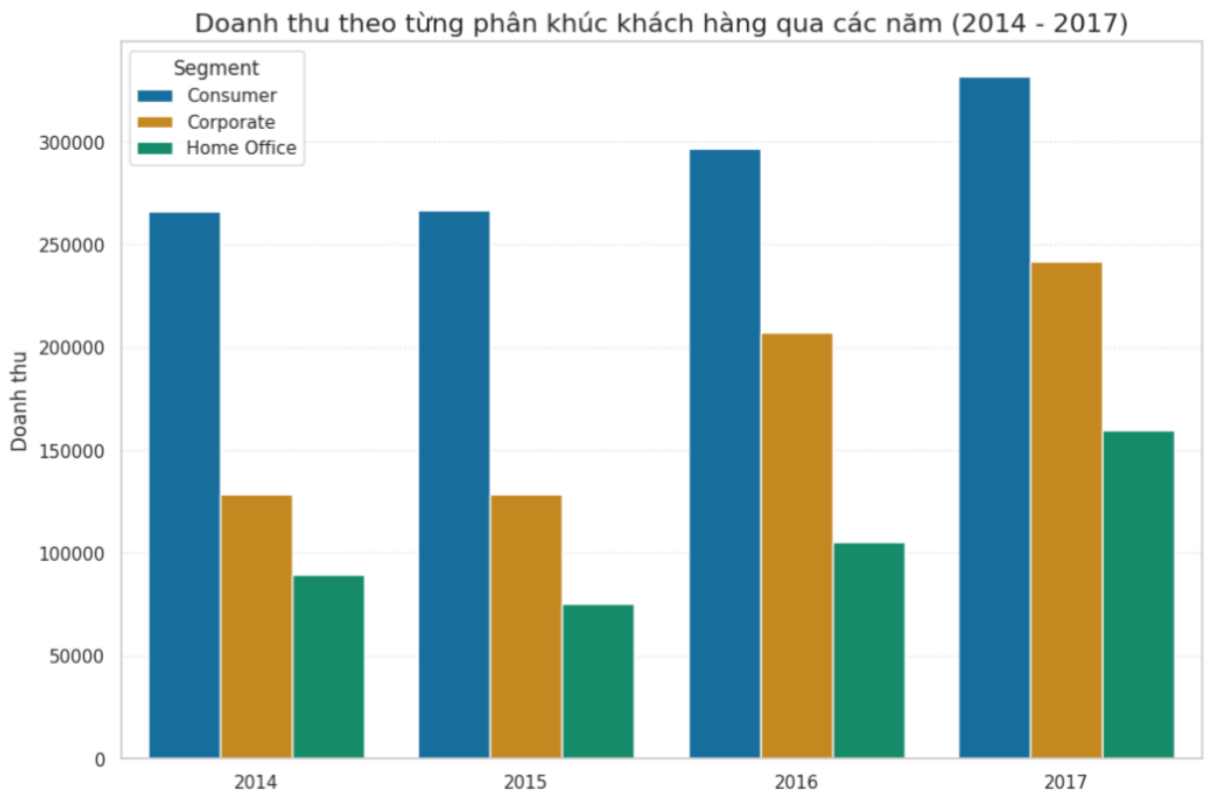
3.1. Biểu đồ thể hiện tổng số

3.1.1. Top 20 bang đóng góp doanh thu nhiều nhất cho cửa hàng



Nhận xét: Quan sát biểu đồ ta thấy được California là bang có tổng doanh thu cao nhất, xếp thứ hai ngay phía sau là New York và đứng thứ ba là Texas. Từ bang thứ 6 là Florida ta thấy được doanh thu giữa các bang đã không có sự chênh lệch quá lớn và gần về cuối ta thấy được mức chênh lệch này là không đáng kể. Từ đó ta thấy được cửa hàng trải rộng thị trường trên khắp nước Mỹ nhưng chỉ tập trung ở một số thị trường nhất định.

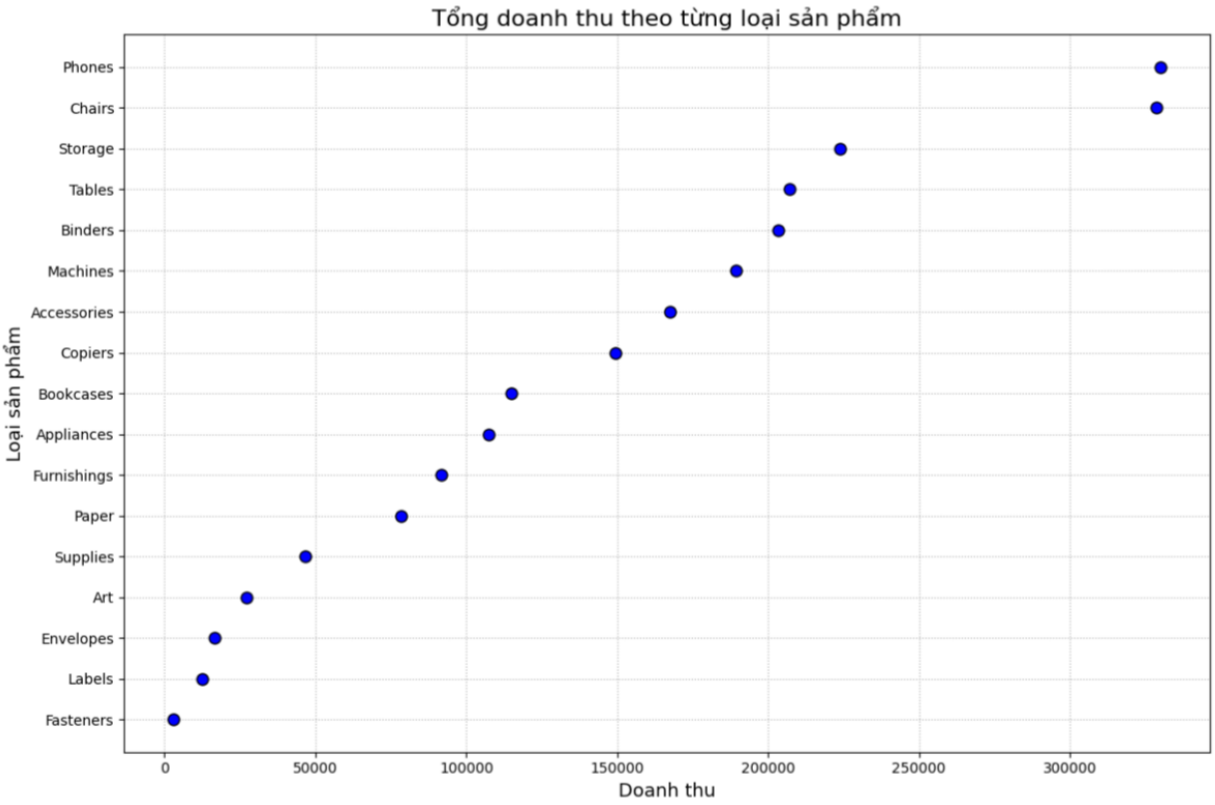
3.1.2. Đóng góp doanh thu của từng phân khúc qua các năm



Nhận xét: Có thể nhận thấy sự chênh lệch giữa các phân khúc khách hàng trên biểu đồ. Đặc biệt, Consumer là phân khúc mang lại nhiều doanh thu nhất cho cửa hàng, tiếp theo sau là Corporate và xếp cuối cùng là Home Office.

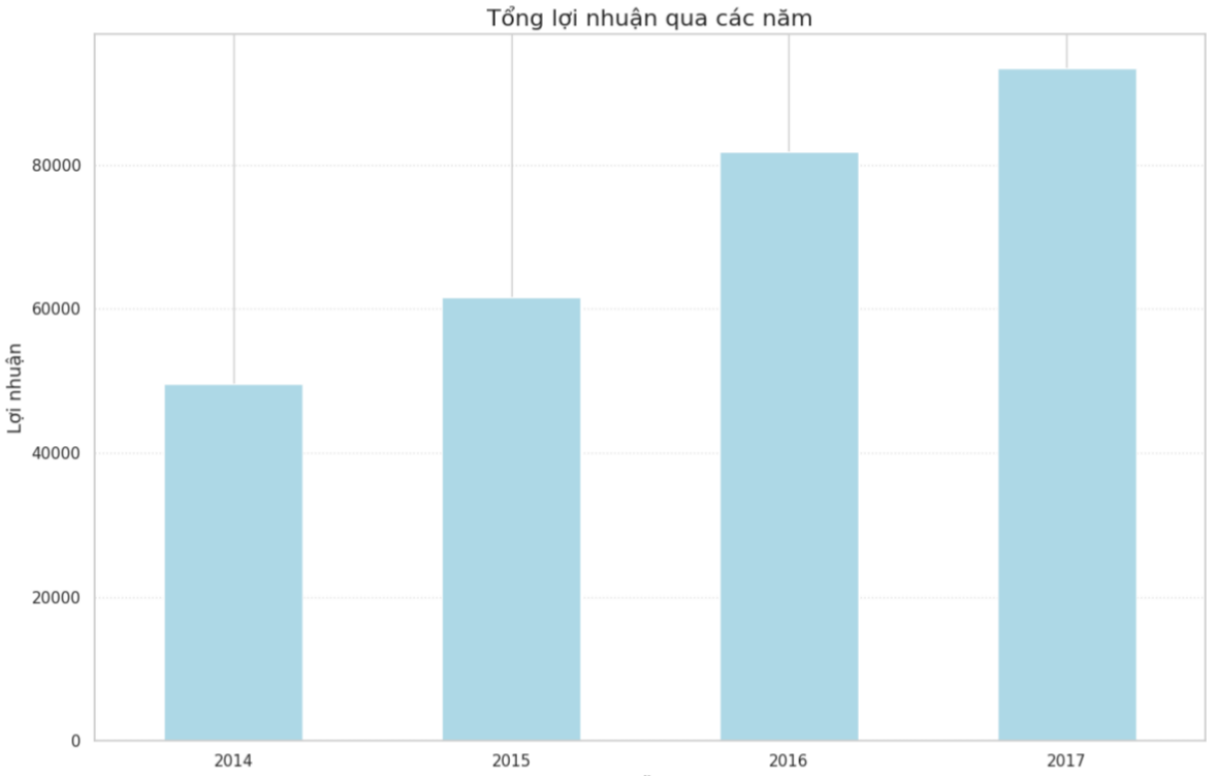
Doanh thu của cửa hàng có xu hướng tăng qua từng năm. Năm 2017 là năm có tổng doanh thu cao nhất, và năm 2016 đứng ngay phía sau. Tuy nhiên, năm 2015 chứng kiến một sự giảm nhẹ trong doanh thu của phân khúc 'Home Office' so với năm 2014.

3.1.3. Doanh thu của cửa theo chi tiết từng loại hình sản phẩm



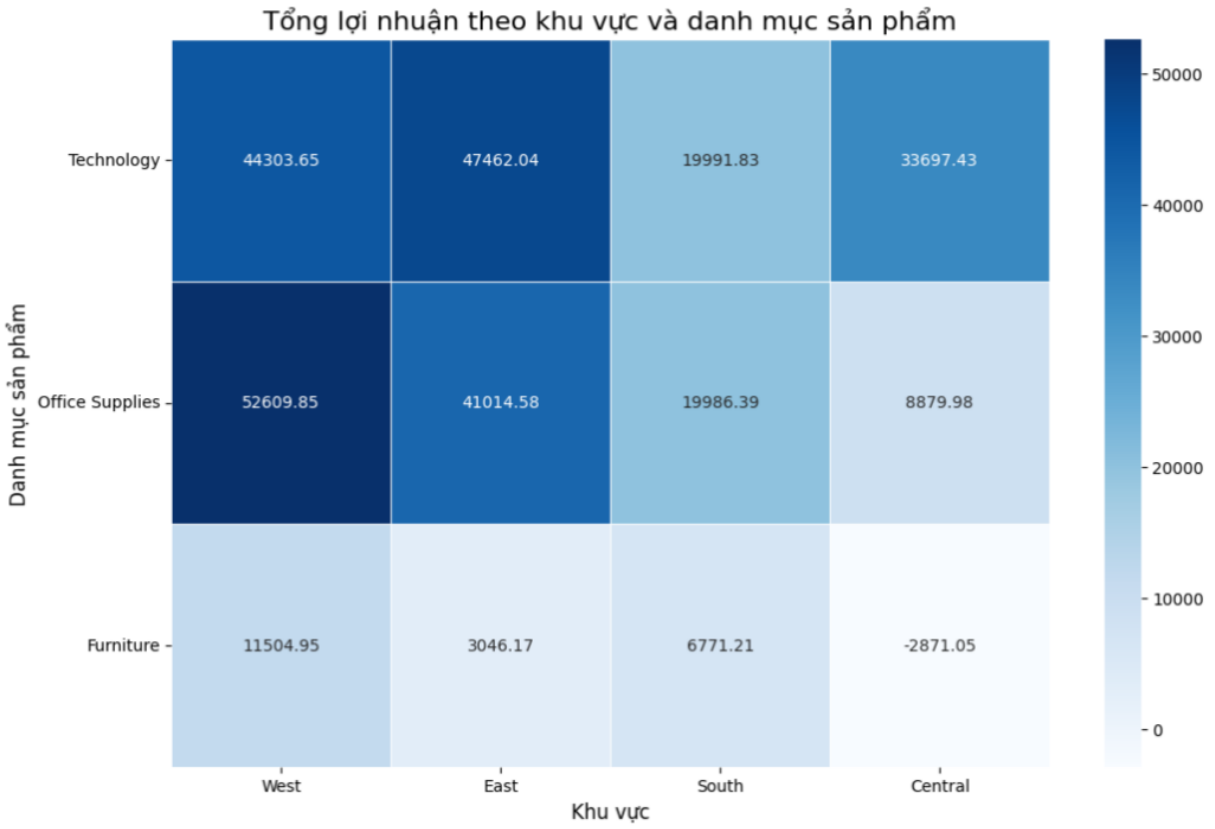
Nhận xét: Phones và Chairs là 2 loại hình sản phẩm đóng vai trò quan trọng trong việc tạo ra doanh thu cho cửa hàng, trở thành những sản phẩm bán chạy nhất. Ngược lại, Fasteners, Labels và Envelopes là những loại sản phẩm ít đóng góp vào doanh thu nhất, trong đó Fasteners là ít được ưa chuộng nhất.

3.1.4. Lợi nhuận của cửa hàng qua các năm



Nhận xét: Năm 2017 ghi nhận mức lợi nhuận cao nhất trong các năm 2014 – 2017, trong khi năm 2014 là năm có mức lợi nhuận thấp nhất. Nhìn chung, lợi nhuận của cửa hàng có xu hướng tăng trưởng từ 2014-2017. Điều này minh chứng cho sự phát triển ngày càng tăng của cửa hàng.

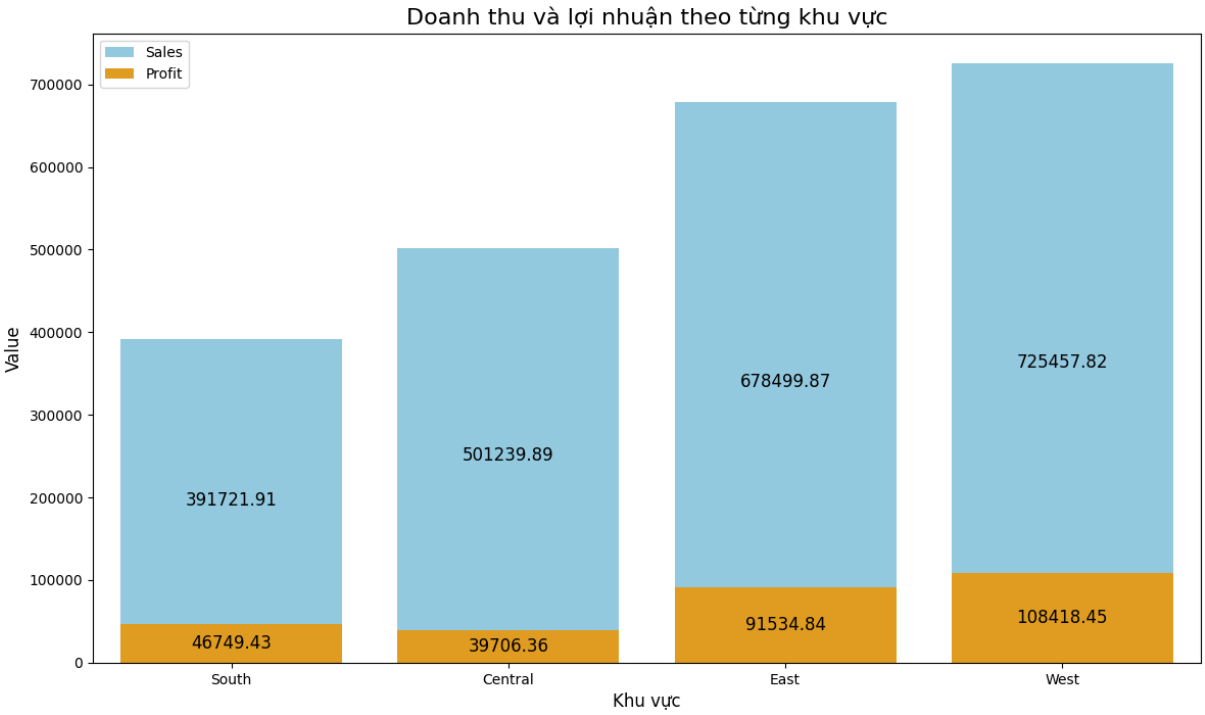
3.1.5. Lợi nhuận của cửa hàng theo từng khu vực và danh mục sản phẩm



Nhận xét: Có thể thấy rằng khu vực phía Tây dẫn đầu về lợi nhuận của cửa hàng, trong khi khu vực Trung tâm lại có lợi nhuận thấp nhất. Trong các danh mục sản phẩm thì Technology có lợi nhuận cao nhất, tiếp theo là Office Supplies, ngược lại Furniture là mục sản phẩm ít đóng góp vào lợi nhuận của cửa hàng nhất.

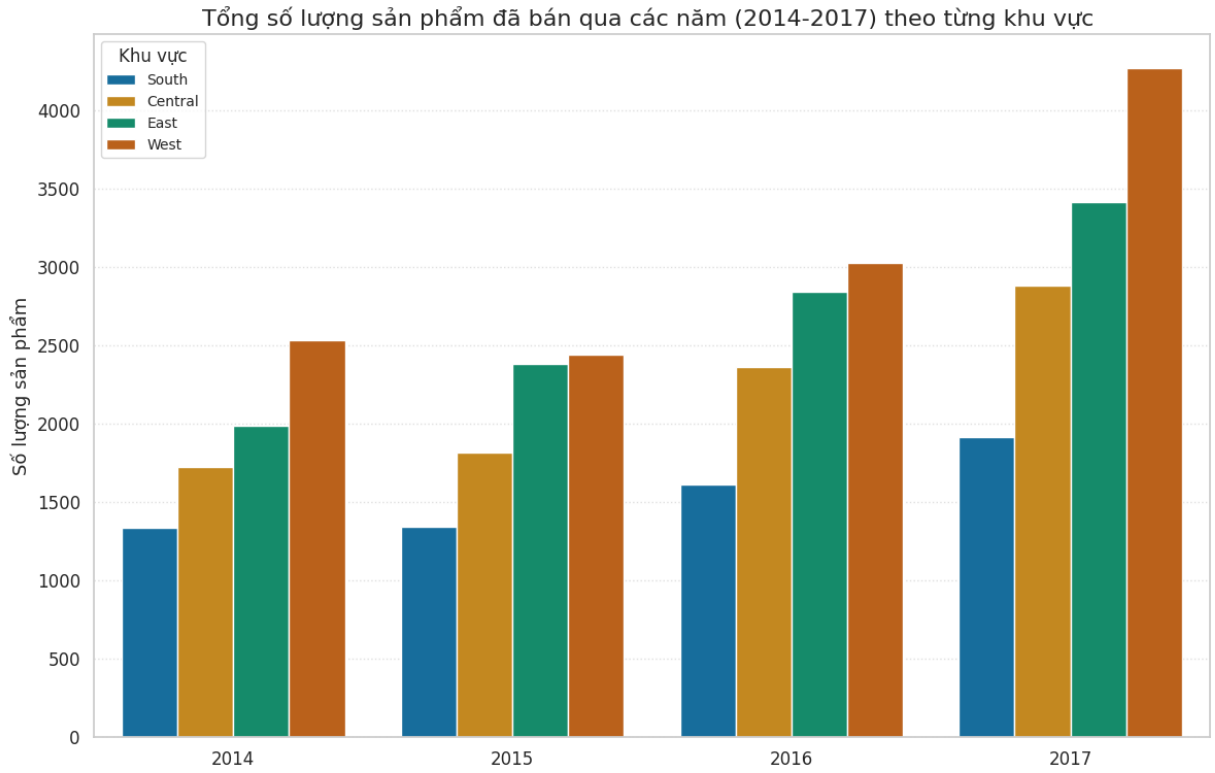
Đặc biệt, danh mục Office Supplies ở khu vực phía Tây đã đóng góp cho cửa hàng một mức lợi nhuận đáng kể, cao hơn hẳn so với các khu vực và danh mục sản phẩm khác. Điều này cho thấy sự phổ biến và sức hút mạnh mẽ của các sản phẩm Office Supplies tại khu vực phía Tây. Trong khi đó, tại khu vực Trung tâm, Furniture lại không tạo ra lợi nhuận cho cửa hàng.

3.1.6. Doanh thu và lợi nhuận của cửa hàng theo khu vực



Nhận xét: Khu vực phía Tây và phía Đông là những khu vực đóng góp vào doanh thu và lợi nhuận cao nhất cho cửa hàng. Trong khi đó, ta thấy được một bất cập ở khu vực Central khi doanh số ở khu vực này cao hơn South nhưng lợi nhuận lại thấp hơn, giải thích cho điều này có thể cửa hàng đã thiết lập tỉ lệ chiết khấu cho khu vực này cao hơn khu vực khác điều này dẫn đến doanh số ở khu vực này lớn nhưng lợi nhuận lại không bằng những khu vực khác.

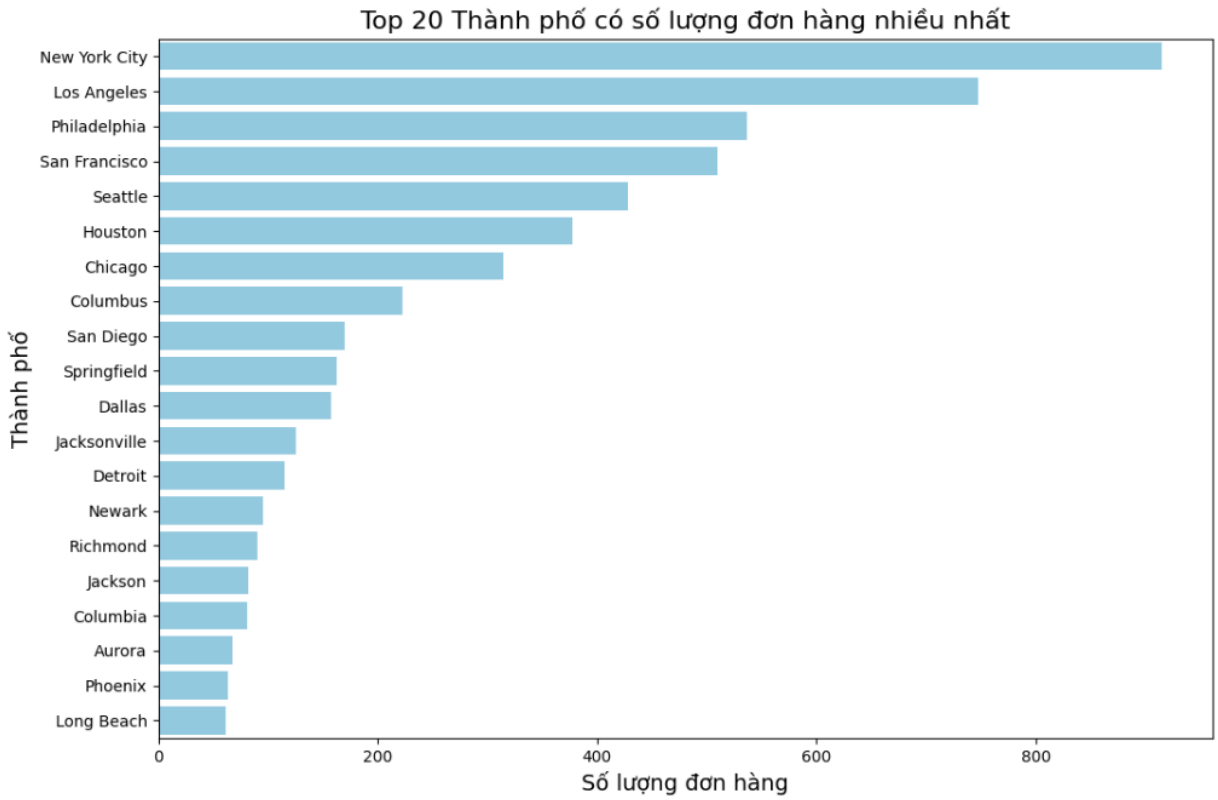
3.1.7. Tổng số lượng sản phẩm đã bán của cửa hàng qua các năm theo từng khu vực



Nhận xét: Ta thấy được số lượng sản phẩm bán được từ 2014 đến cuối 2017 đã có sự tăng trưởng rõ rệt. Với số lượng dẫn đầu theo thứ tự là West – East – Central – South.

Tuy nhiên trong giai đoạn năm 2015 ta thấy được sự suy giảm nhẹ về sản lượng ở các khu vực, ngoại trừ khu vực East đã có sự tăng trưởng ổn định qua các năm.

3.1.8. Top 20 thành phố có số lượng đơn hàng nhiều nhất



Nhận xét: Có thể thấy, New York City là thành phố đứng đầu với số lượng sản phẩm bán ra nhiều nhất, tiếp theo phía sau là thành phố Los Angeles, đứng thứ ba là thành phố Philadelphia. Trong khi đó, Long Beach và Phoenix là hai thành phố có số lượng sản phẩm bán ra có phần kém hơn so với các thành phố khác trong bảng xếp hạng 20 thành phố có số lượng sản phẩm bán chạy nhất của cửa hàng.

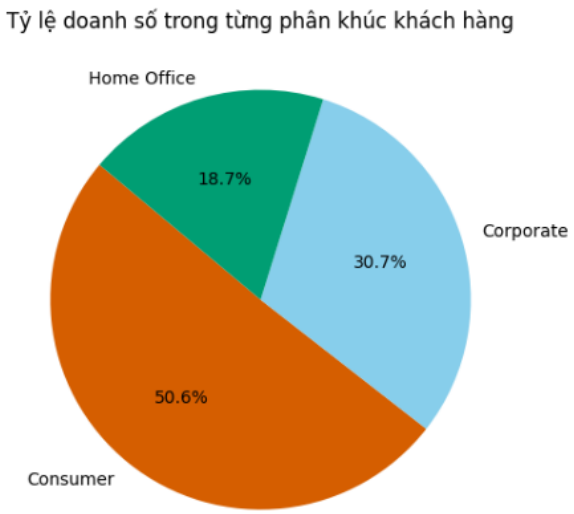
Ngoài ra ta thấy được sự ngược lại ở biểu đồ doanh thu theo các bang trước trong khi bang Califonia là bang có doanh thu cao nhất và New York chỉ xếp thứ 2 thì Los Angeles (thành phố ở California) lại bán được ít hàng hơn New York giải thích cho điều này thì do California có diện tích gấp khoảng 3 lần so với New York ngoài ra ta cũng có thể thấy trong top 20 ngoài Los Angeles còn San Francisco, San Diego, Newark, Richmond và Long Beach là các thành phố thuộc bang California.

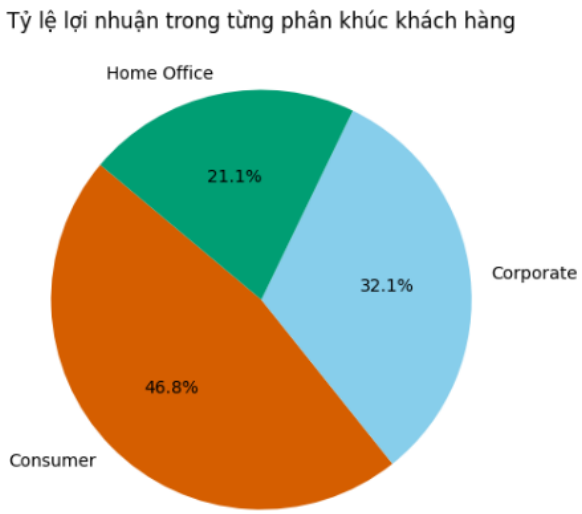
3.2. Biểu đồ thể hiện tỉ lệ

3.2.1. Tỉ lệ đơn hàng trong từng danh mục sản phẩm

Nhận xét: Theo biểu đồ, danh mục sản phẩm chiếm tỷ lệ cao nhất trong tổng số các đơn hàng là văn phòng phẩm, với 53,1% chiếm quá nửa tổng số đơn hàng của cả cửa hàng. Tiếp theo là nhóm sản phẩm nội thất, với 25,0% và công nghệ chiếm tỷ lệ thấp nhất, với 21,9%.

3.2.2. Tỉ lệ doanh số và tỷ lệ lợi nhuận trong từng phân khúc khách hàng





Nhận xét: Theo biểu đồ, phân khúc khách hàng cá nhân chiếm tỷ lệ doanh thu và lợi nhuận cao nhất, lần lượt là 50,6% và 46,8%. Xếp thứ hai là khách hàng doanh nghiệp với doanh thu là 30,7% và lợi nhuận chiếm 32,1%. Tỷ lệ doanh thu và lợi nhuận thấp nhất là ở phân khúc văn phòng tại nhà.

Ngoài ra, ta thấy được so với doanh số thì lợi nhuận ở nhóm khách hàng cá nhân đã có sự suy giảm. Giải thích cho ý này thì có thể do cửa hàng có nhiều ưu đãi chiết khấu cao hơn đối với nhóm khách hàng này điều này dẫn đến doanh thu không tương xứng với lợi nhuận.

3.2.3. Tỉ lệ đơn hàng của từng loại hình sản phẩm trong mỗi danh mục

Nhận xét: Theo biểu đồ, tỉ lệ giữa 3 danh mục tương tự với nhận xét đã đề cập ở biểu đồ trước tuy nhiên ở đây ta thấy thêm được tỉ lệ của từng loại hình sản phẩm trong từng danh mục.

Cụ thể, Binder và Paper là hai mặt hàng được mua nhiều nhất trong danh mục Office Supplies, 2 nhóm sản phẩm đã chiếm gần một nửa danh mục và chiếm gần 1/3 tổng các đơn hàng. Furnishings và Phones lần lượt là mặt hàng được mua nhiều nhất trong danh mục Furniture và Technology chúng cũng chiếm tỉ lệ ngang nhau khoảng một nửa của từng danh mục tương ứng cũng là hơn 1/10 của tổng thể các đơn hàng (một nửa của 1/5).

3.2.4. Tỉ lệ doanh thu các danh mục

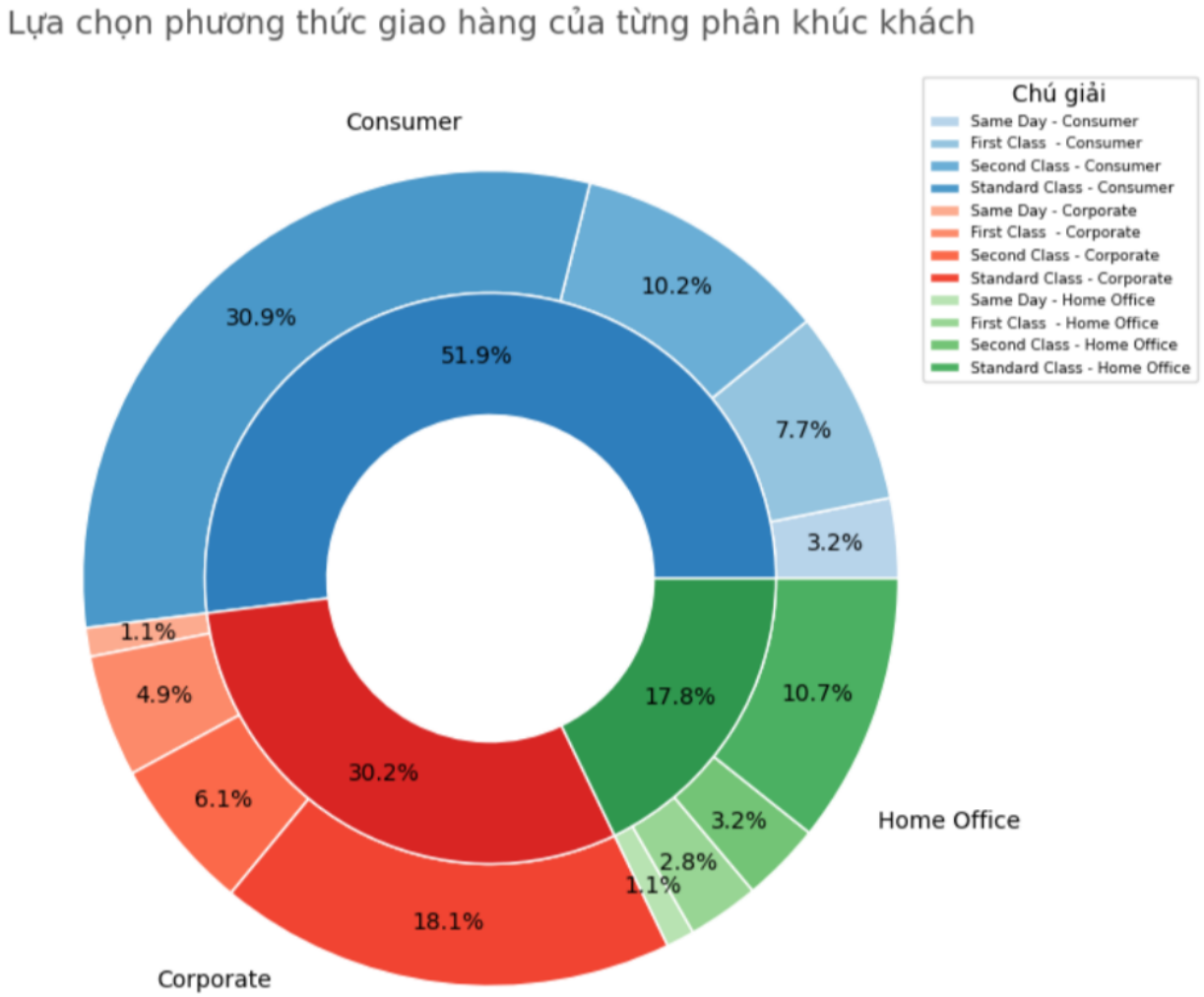
Doanh thu các danh mục sản phẩm theo từng khu vực

Nhận xét: Theo biểu đồ, ta thấy doanh thu của 2 khu vực West và East không chênh lệch quá lớn, đều đóng góp lượng lớn doanh thu cho cửa hàng chiếm gần 1/3 tổng doanh thu.

Furniture là danh mục mang lại doanh thu cao nhất ở khu vực West, trong khi đó ở 3 khu vực còn lại thì Technology có doanh thu cao nhất. Office supplies là danh mục có doanh thu thấp nhất ở khu vực West và East, còn ở Central và South thì doanh thu thấp nhất lại đến từ Furniture.

Tuy vậy ta thấy được ở cả 4 khu vực tỷ lệ doanh thu giữa các danh mục sản phẩm trong cùng một khu vực không có sự chênh lệch quá lớn, có thể thấy ở từng khu vực nhu cầu với các danh mục sản phẩm là gần như nhau.

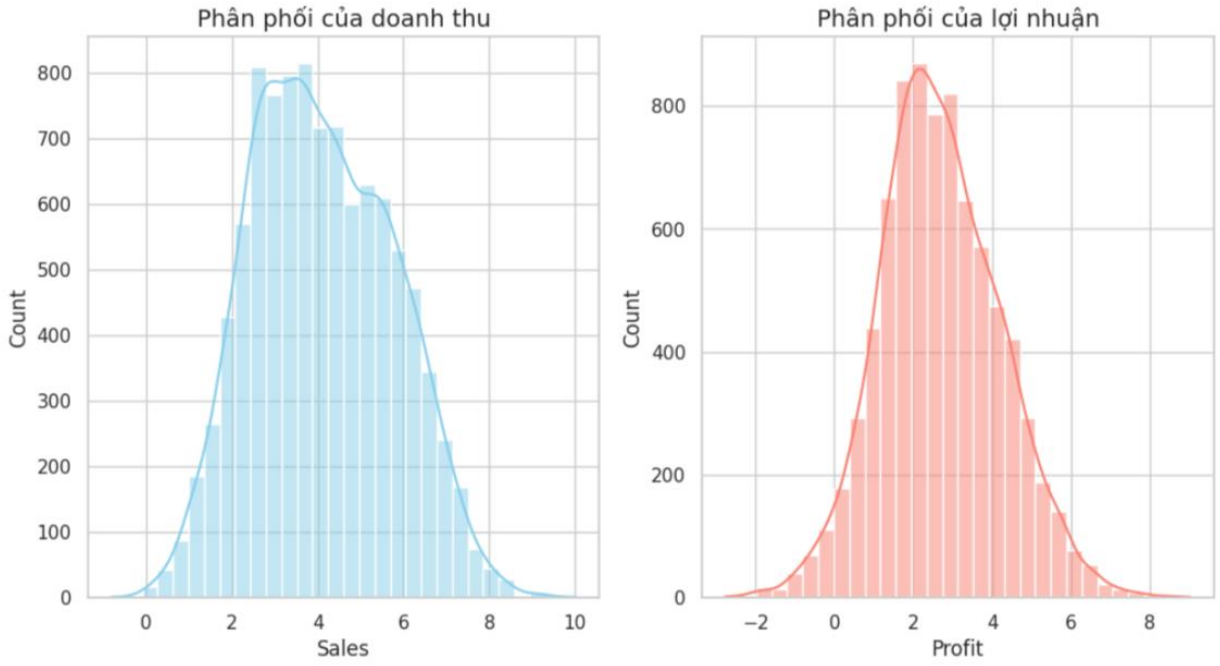
3.2.5. Tỉ lệ lựa phương thức giao hàng trong mỗi phân khúc khách hàng



Nhận xét: Ở cả 3 phân khúc khách hàng, phương thức vận chuyển Standard Class chiếm tỷ lệ cao nhất so với 3 phương thức còn lại. Tiếp theo thứ tự thì cả 3 nhóm khách đều lựa chọn Second Class, First Class, Same Day. Giải thích cho ý này là do theo thứ tự trên thì chi phí cho các phương thức giao hàng này cũng tăng dần, cũng đồng thời chất lượng dịch vụ cũng sẽ tốt hơn. Do vậy khách hàng thường ưu tiên các phương thức có chi phí từ thấp đến cao.

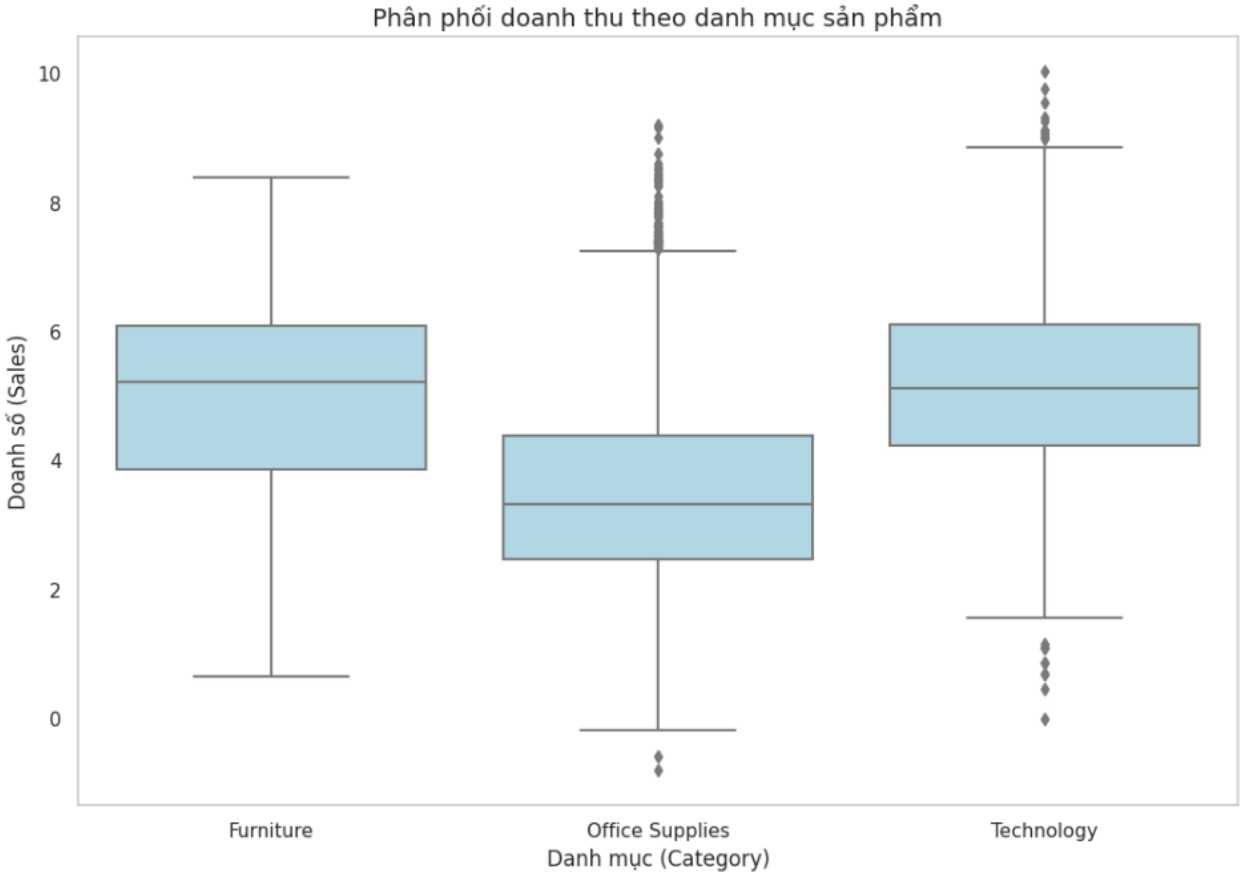
3.3. Biểu đồ thể hiện phân phối

3.3.1. Phân phối của doanh thu và lợi nhuận



Nhận xét: Biểu đồ này cho thấy rằng phân phối doanh thu và lợi nhuận có hình dạng tương tự nhau, nhưng phân phối lợi nhuận có xu hướng tập trung về trung tâm hơn còn phân phối doanh thu thì có vẻ đang lệch phải một ít.

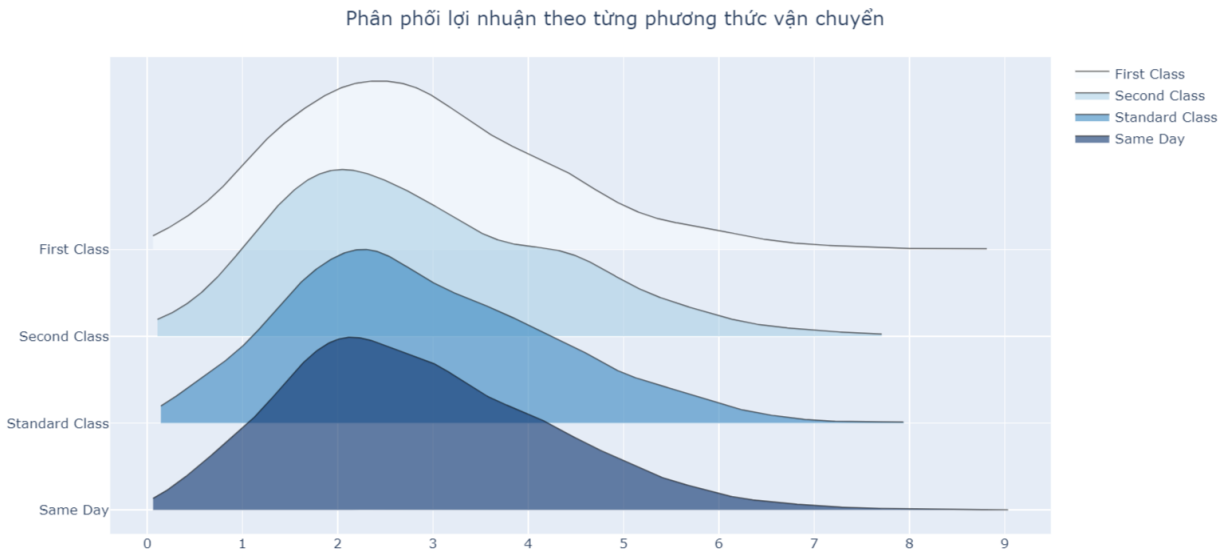
3.3.2. Phân phối của doanh thu theo từng danh mục sản phẩm



Nhận xét: Ta nhận thấy doanh thu ở Office Supplies có xu hướng tập trung ở mức thấp hơn bởi vì các loại văn phòng phẩm như bút, kẹp giấy, mực, … thường rẻ hơn nhiều so với các sản phẩm khác điều này cũng dẫn đến doanh thu ở nhóm này cũng thường tập trung ở mức thấp hơn so với 2 nhóm khác.

Ngoài ra ta cũng thấy số lượng outlier ở phần trên của danh mục này rất lớn, điều này bởi lẽ ta biết đại đa số văn phòng phẩm có giá trị thấp nhưng cũng có một số món hàng có giá trị khá cao như máy đóng sách, máy mở thư tự động, … chính những đơn hàng với những sản phẩm như vậy tạo nên số lượng nhiễu lớn ở phía trên boxplot ở danh mục này.

3.3.3. Phân phối lợi nhuận theo từng phương thức vận chuyển

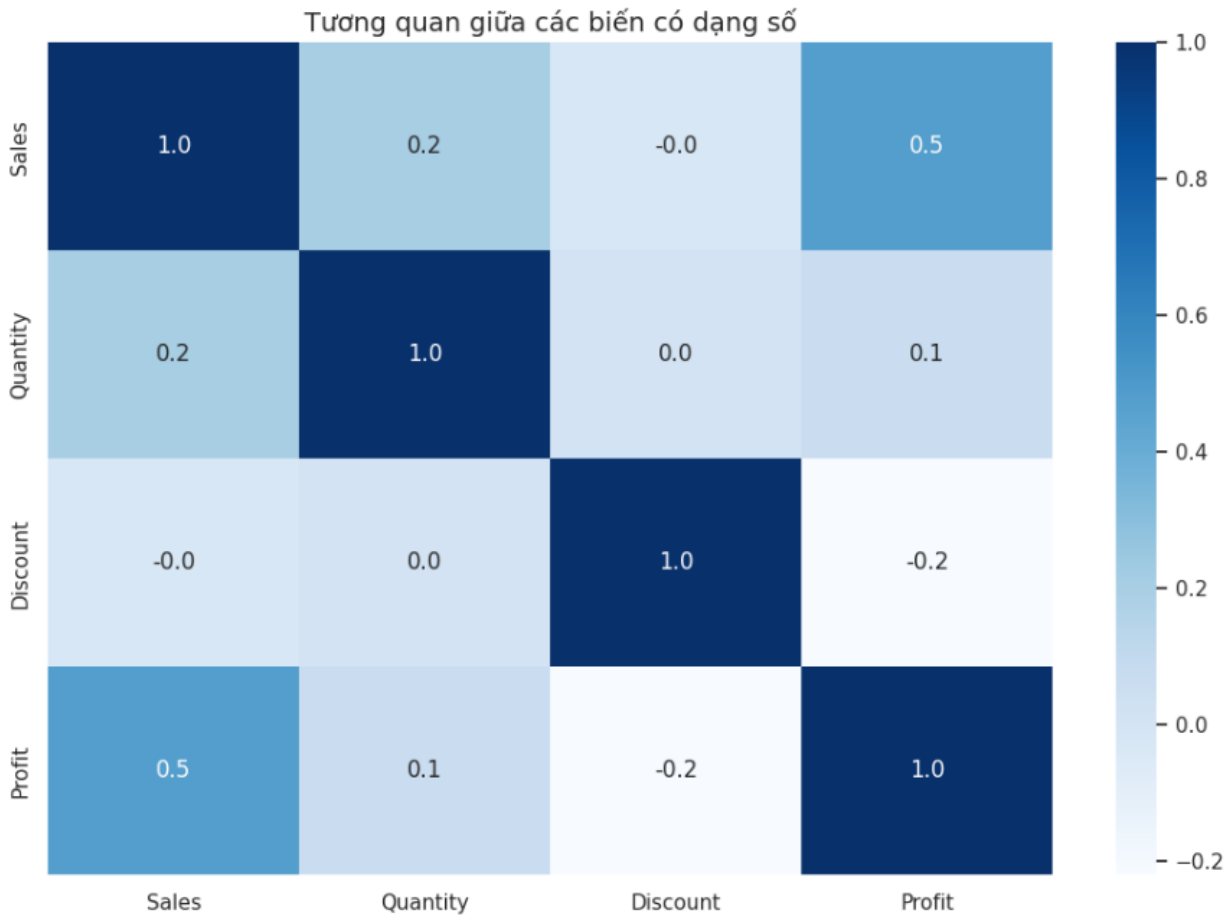


Nhận xét: Có thể thấy, hình dạng phân phối của cả bốn phương thức tương đồng nhau, tuy nhiên so với biểu đồ phân phối lợi nhuận tổng thể thì chúng lệch phải khá nhiều.

Ngoài ra, ta thấy được ở hai phương thức vận chuyển là First Class và Same Day, phổ giá trị trải rộng hơn so với hai phương thức còn lại. Điều này được lý giải vì First Class và Same Day là phương thức vận chuyển cao cấp hơn, nên chi phí khách bỏ ra là nhiều hơn và ta nhận được lợi nhuận cũng là cao hơn.

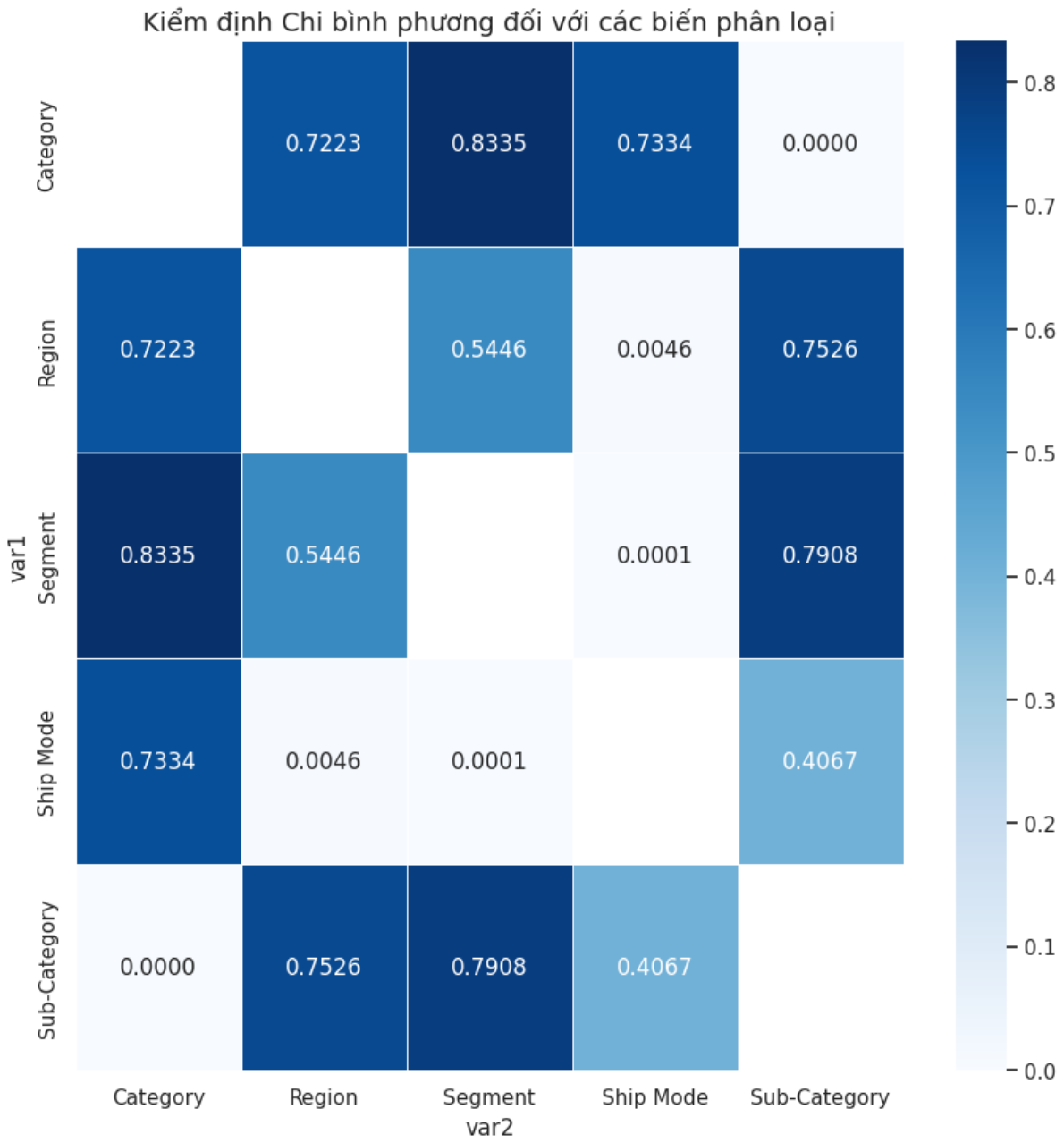
3.4. Biểu đồ thể hiện sự tương quan

3.4.1. Tương quan giữa các biến số



Nhận xét: Đây là heatmap với các giá trị bên trong là chỉ số Pearson để kiểm tra độ tương quan của từng cặp biến, ta thấy được ngoại trừ giữa Sales và Profit có một sự tương quan thuận tương đối mạnh thì tất cả các biến còn lại đều có sự tương quan rất yếu hoặc là không có.

3.4.2. Tương quan giữa các biến phân loại



Dựa trên bảng kiểm định chi bình phương đối với các biến phân loại, bên cạnh đó đặt ra các giả thuyết để đưa ra nhận xét về mối quan hệ giữa các biến này. Giả thuyết của chúng ta như sau:

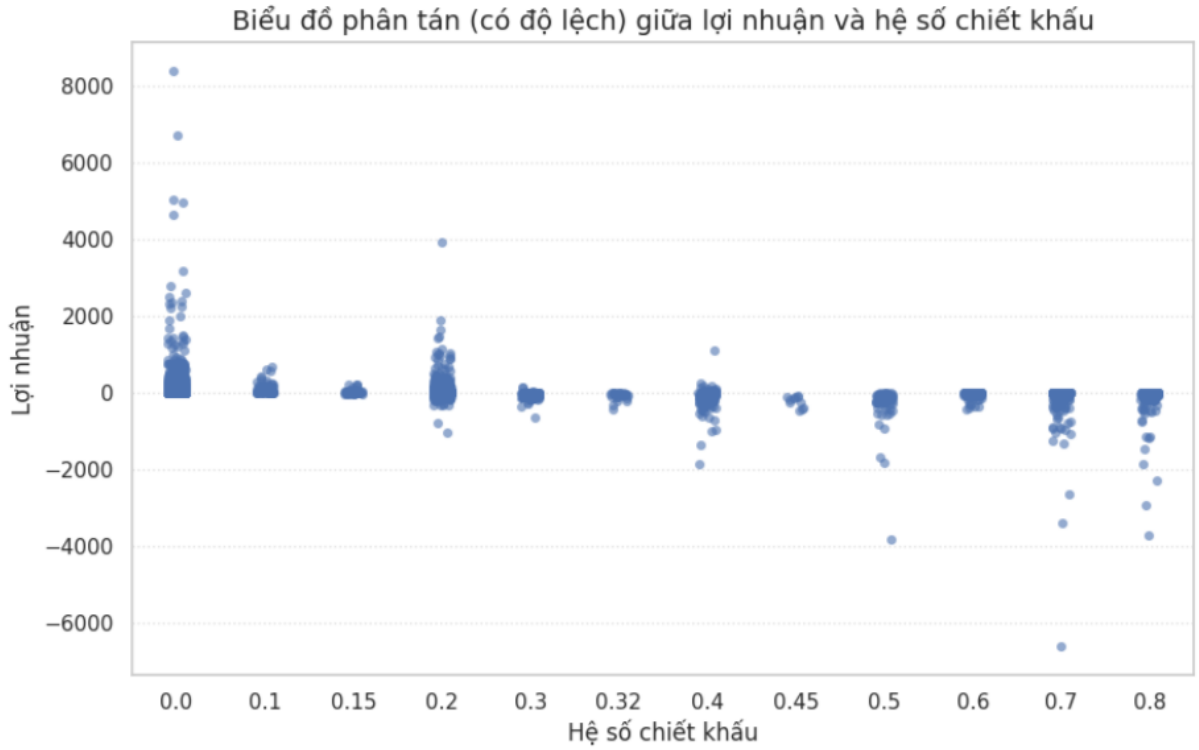
- H0: [Biến 1] và [Biến 2] là 2 biến độc lập.
- Ha: [Biến 1] và [Biến 2] là 2 biến không độc lập.

Và với khoảng tin cậy 95%, ta có thể đưa ra các nhận xét sau:

Dựa vào bảng trên, ta có thể thấy rằng các biến danh mục sản phẩm và loại hình sản phẩm là phụ thuộc nhau nhưng lại độc lập với các biến khác.

- Biến phương thức vận chuyển độc lập với các biến liên quan đến loại hình sản phẩm nhưng không độc lập với phân khúc khách hàng và khu vực.
- Tương tự như vậy với biến phân khúc khách hàng, biến này độc lập với các biến khác chỉ ngoại trừ biến phương thức vận chuyển. Điều này tuy không ám chỉ rằng có mối quan hệ nhân quả giữa 2 biến, nhưng ta có thể đặt ra giả thuyết rằng các phân khúc khách hàng khác nhau sẽ thường chuộng các phương thức vận chuyển khác nhau.
- Cuối cùng, tương tự với biến khu vực, biến này độc lập với các biến khác nhưng khi kiểm định thì lại không độc lập với biến phương thức vận chuyển. Điều này cũng không ám chỉ rằng tồn tại mối quan hệ nhân quả nào đó giữa biến khu vực và phương thức giao hàng. Nhưng ta có thể đặt ra giả thuyết rằng có một số khu vực thường ưu tiên các phương thức giao hàng khác nhau.

3.4.3. Tương quan giữa lợi nhuận và tỉ lệ chiết khấu



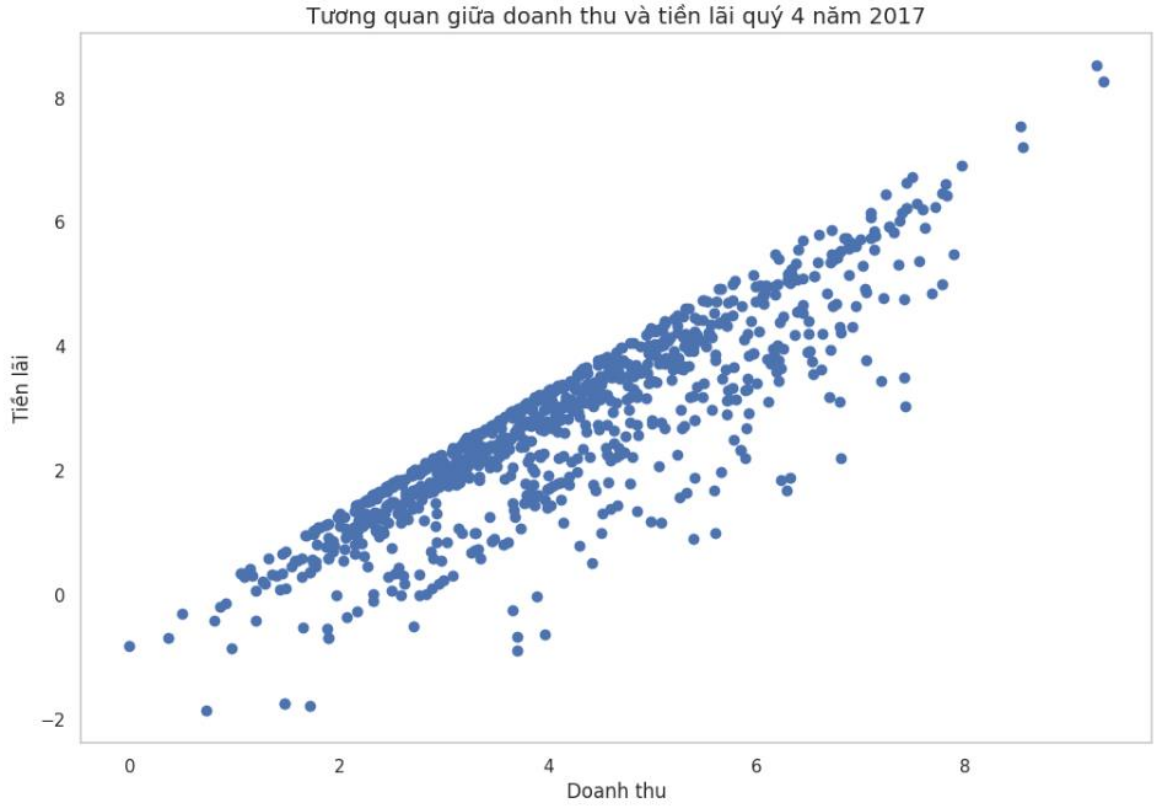
Nhận xét: Như đã thể hiện ở hệ số tương quan giữa 2 biến lợi nhuận và biến hệ số chiết khấu, ta có thể thấy rằng với một hệ số tương quan âm, 2 biến này có mối tương quan nghịch chiều, và hình vẽ trên đã củng cố kết luận đó. Ta thấy được kể từ mức chiết khấu

0.2 đã có sự những giao dịch mang lại lợi nhuận âm và khi mức chiết khấu tăng dần thì các giao dịch mang lợi nhuận dương càng ít và lợi nhuận âm càng nhiều.

Tuy nhiên, ta thấy được chỉ số Pearson đã tính trước đó (-0.2) đã phản ánh đúng mối tương quan nghịch giữa 2 biến nhưng vì số trường hợp như vậy khá ít so với tổng thể điều này dẫn đến mối tương quan này biểu hiện rất yếu.

3.4.4. Tương quan giữa tiền lãi và doanh thu trong quý 4 năm 2017

Để có thể trình bày biểu đồ một cách tốt nhất, ở đây ta sẽ chỉ quan sát phần tiền lãi và doanh thu vào quý 4 năm 2017 và ta sẽ biến đổi log cho cả 2 cột.



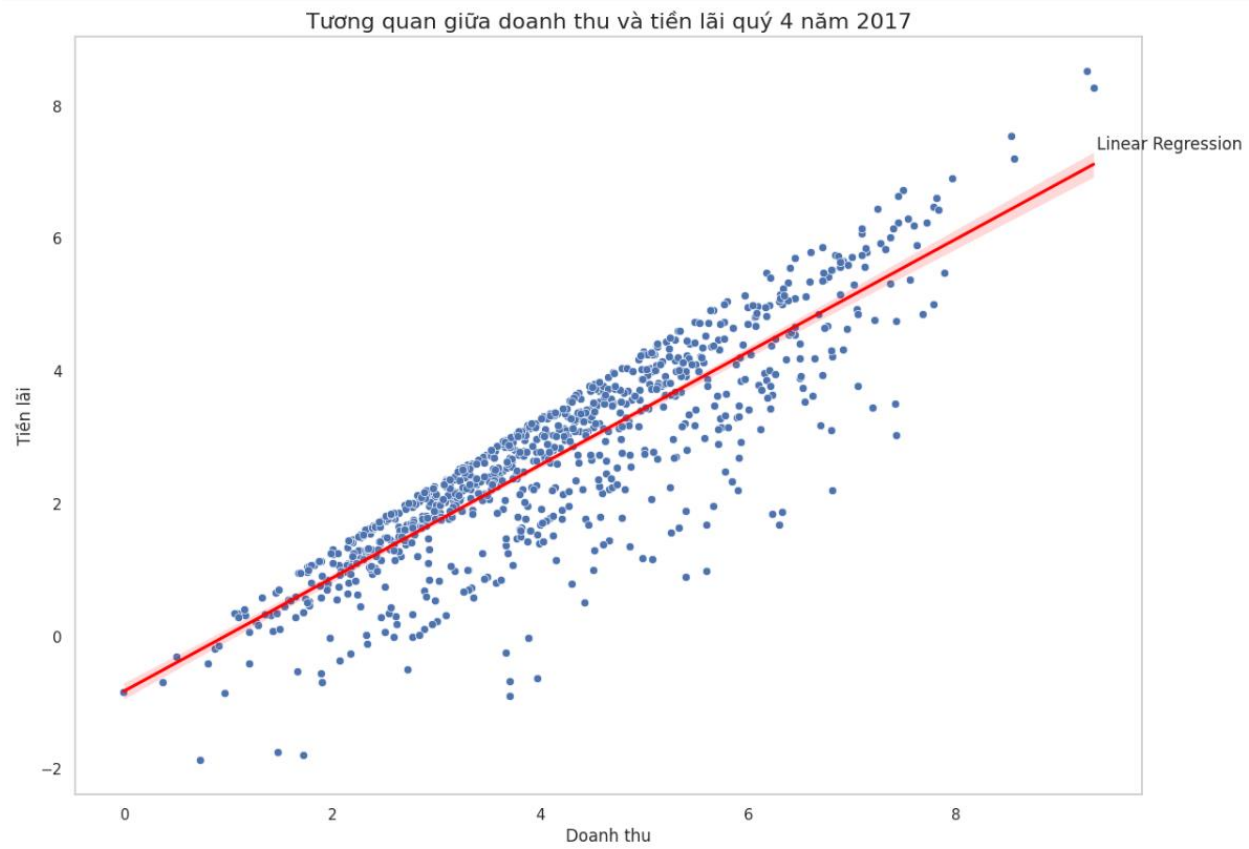
Nhận xét: Ta thấy được sự tương quan thuận khá mạnh giữa tiền lãi và doanh thu ta có thể kết luận rằng khi doanh thu càng cao thì tiền lãi cửa hàng nhận được càng nhiều. Khác với giá trị của chỉ số pearson thu được từ heatmap trên (0.5) ở đây sau khi biến đổi ta có được một giá trị hoàn toàn khác.

```
Chỉ số Pearson giữa Profit và Sales: 0.8785875795998868
```

Tuy nhiên để chắc chắn với kết luận trên ta sẽ dùng một phương pháp khác để kiểm chứng kết luận một lần nữa.

3.5. Nhóm biểu đồ khác

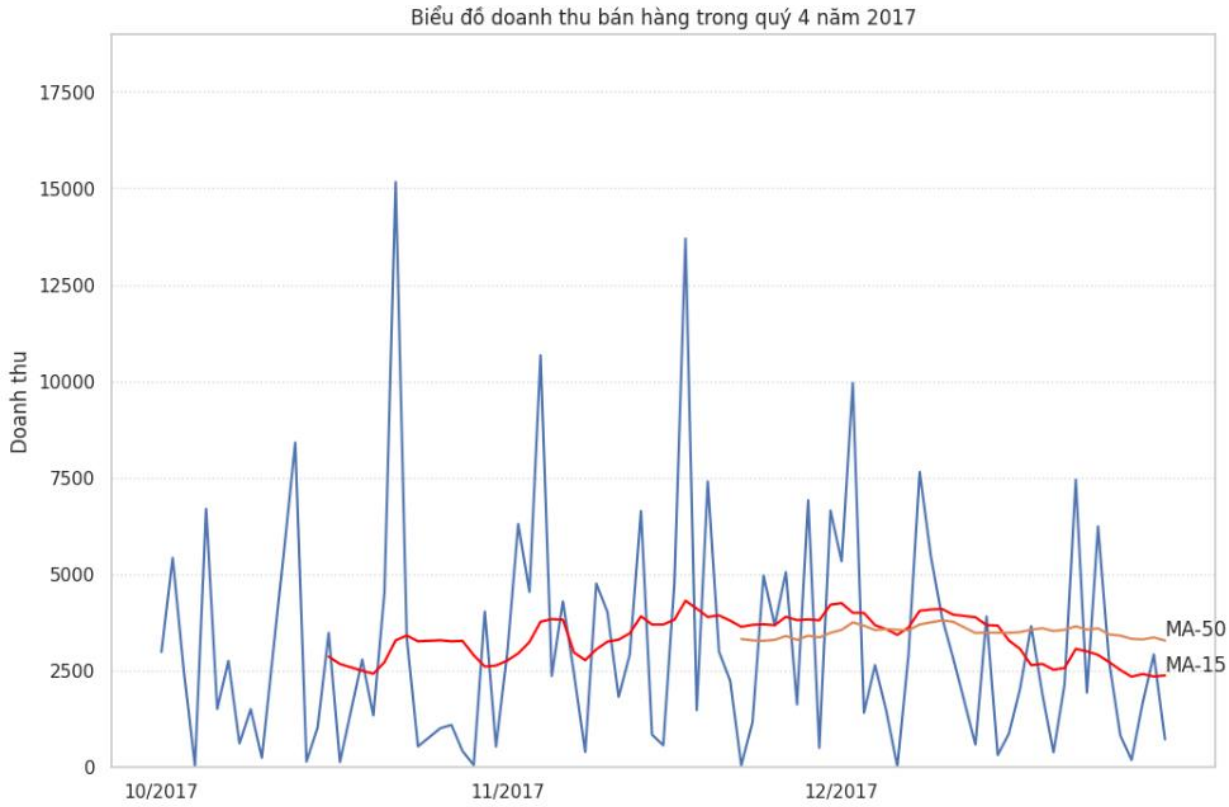
3.5.1. Tương quan giữa tiền lãi và doanh thu trong quý 4 năm 2017 với đường hồi quy tuyến tính và confidence band



Nhận xét: Từ đường hồi quy ta thấy được kết luận về sự tương quan mạnh giữa 2 biến tiền lãi và doanh thu là đúng, đồng thời confidence band hẹp xuất hiện trên đường hồi quy tuyến tính cũng thể hiện đường hồi quy đã xấp xỉ khá chính xác mối quan hệ giữa hai biến này

3.5.2. Doanh thu của cửa hàng trong quý 4 năm 2017

Với kết luận trên ta có thể xem xét doanh thu của cửa hàng trong khoảng thời gian gần đây để xác định tình hình kinh doanh của cửa hàng. Cụ thể, ta sẽ dùng phương pháp làm trơn bằng đường trung bình động cụ thể ở đây là MA-15 và MA-50 để dễ dàng xem xét xu hướng của doanh thu cửa hàng trong quý 4 năm 2017

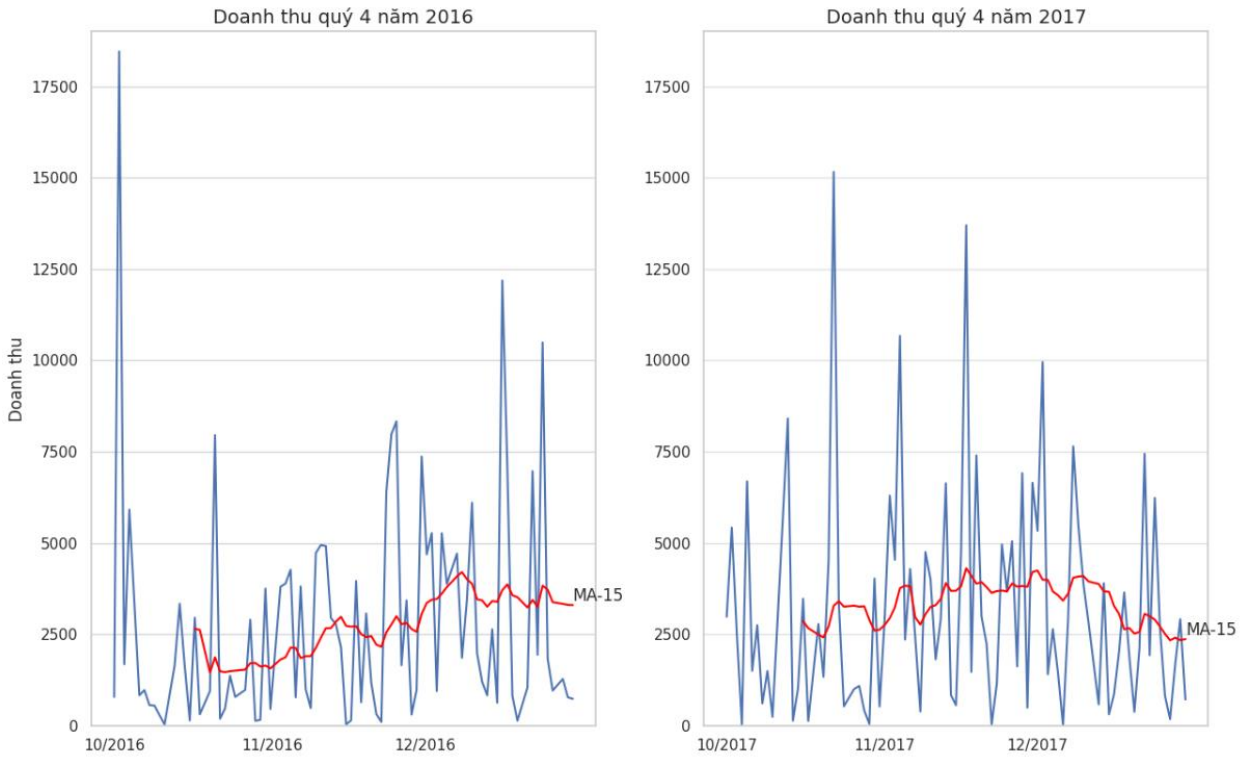


Nhận xét: Trong quý 4 năm 2017, dựa vào đường MA-15 ta thấy được doanh thu mỗi ngày có xu hướng dao động trong khoảng 2500 - 4000. Dù doanh thu có biến động qua từng ngày nhưng biên độ không lớn cho thấy doanh thu của cửa hàng luôn duy trì ở mức ổn định.

Nhiều ngày đạt mức doanh thu với biên độ chênh lệch rất lớn vượt lên MA-15 và MA-50 so, tuy nhiên cũng có những ngày doanh thu trượt xuống dưới 2 đường này nhưng biên độ không quá lớn. Cho thấy doanh thu cửa hàng đang tăng trưởng tốt

Tuy nhiên trong những ngày gần cuối năm khi MA-15 trượt xuống so với MA-50 cho thấy trong giai đoạn này tăng trưởng doanh thu có xu hướng giảm so với những ngày trước đó.

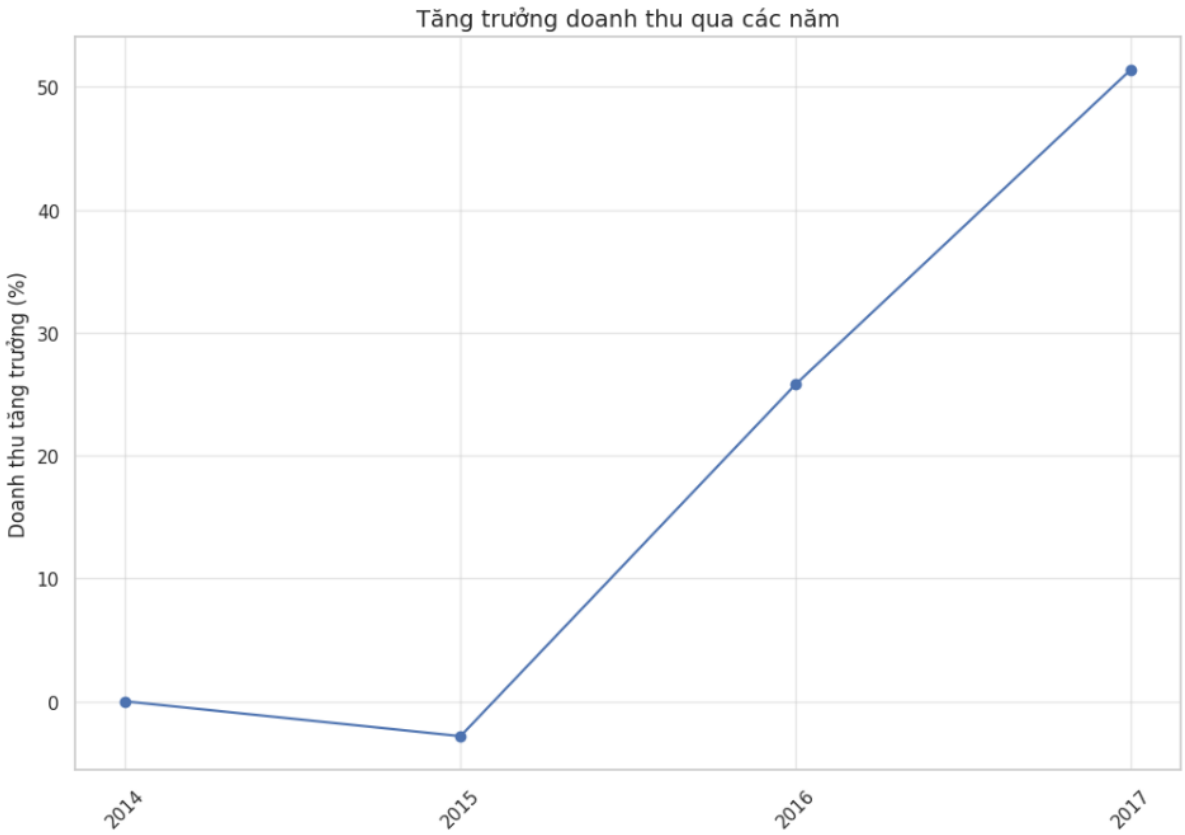
3.5.3. So sánh doanh thu trong quý 4 năm 2017 so với cùng kỳ năm trước



Nhận xét: Cụ thể khi so sánh MA-15 giữa 2 năm ta thấy được doanh thu trong kỳ năm 2016 dao động khoảng từ 1,300 – 4,000 trong khi cũng kỳ này năm 2017 doanh thu dao động trong khoảng từ 2,500 – 4,000.

Ngoài ra dao động trong kỳ năm 2017 ta thấy được doanh thu được duy trì ổn định trong khoảng này trong khi cùng kỳ năm trước thì thường dao động ở gần biên dưới của khoảng. Tuy nhiên, ta cũng thấy được doanh thu ở cả 2 kỳ đều có xu hướng tăng trưởng ổn định, để chắc cho ý này ta sẽ kiểm tra tình hình kinh doanh của cửa hàng qua các năm.

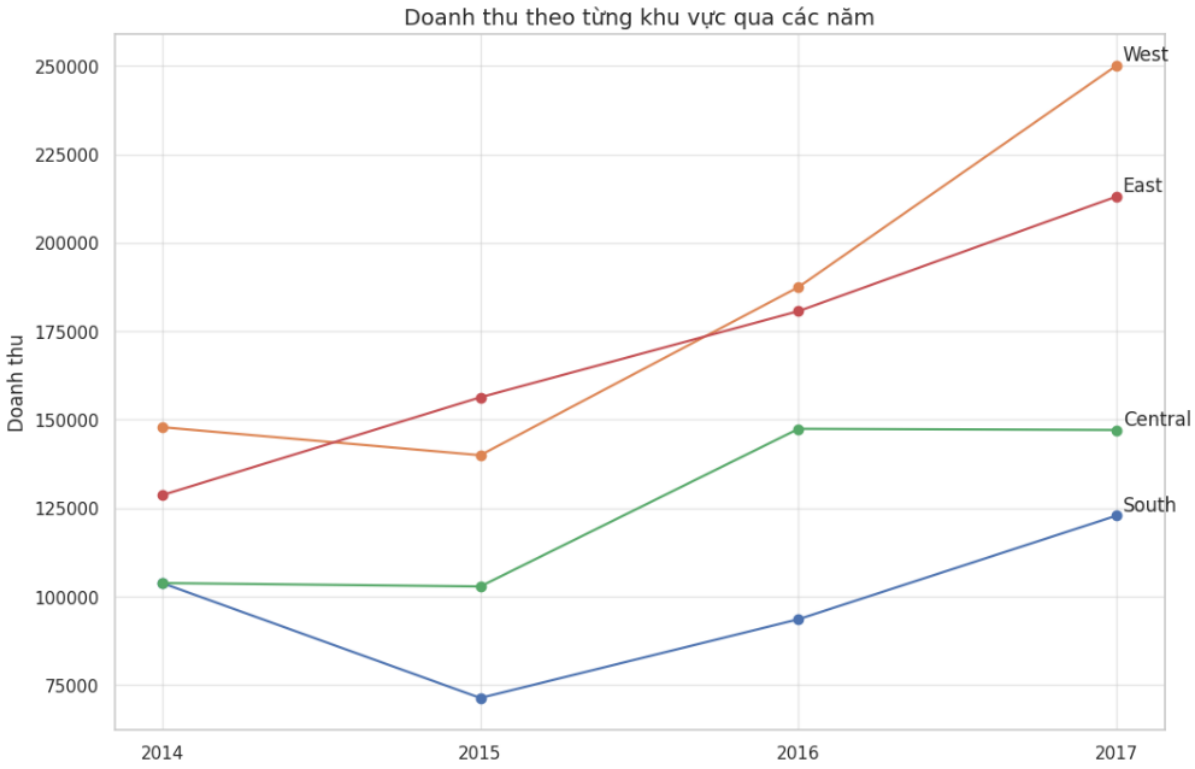
3.5.4. Tăng trưởng doanh thu từ 2014 – 2017



Nhận xét: Như ý trên đã thấy, doanh thu của cửa hàng luôn tăng trưởng ổn định do vậy ta thấy được doanh thu qua các năm đều phát triển rất tốt ngoại trừ một nhịp chỉnh nhẹ trong năm 2015

3.5.5. Doanh thu qua các năm của từng khu vực

Để nắm rõ hơn về biến động doanh thu của cửa hàng ta sẽ tìm hiểu chi tiết về doanh thu của từng khu vực qua các năm.



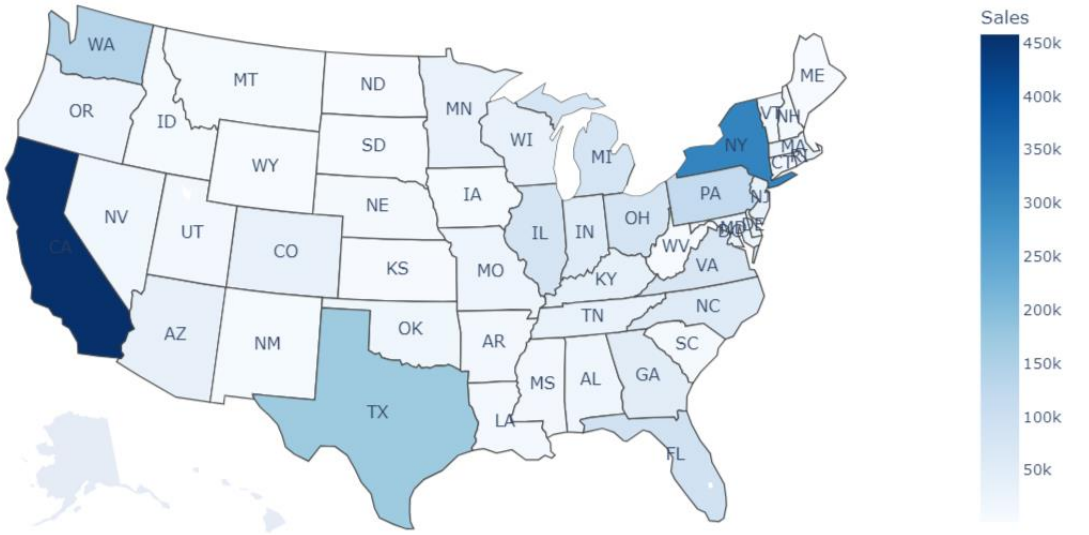
Nhận xét: Đóng góp lớn nhất cho doanh thu của cửa hàng là các khách ở khu vực miền Tây nước Mỹ và thấp nhất là đến từ miền Nam.

Ta thấy được doanh thu các khu vực phát triển đồng pha theo tổng thể ở biểu đồ trên.Tuy nhiên, doanh thu ở khu vực phía Đông vẫn duy trì ổn định qua các năm, không nhận sự biến động giai đoạn năm 2015

Đồng thời, giai đoạn năm 2017 đã có trục trặc gì đó khiến doanh thu của khu vực trung tâm suy giảm, có vẻ khu vực trung tâm cũng là nguyên nhân khiến cho ta thấy doanh thu cửa hàng suy giảm vào những ngày cuối năm trong quý 4 mà ta đã đề cập bên trên

3.5.6. Phân bố doanh thu theo từng tiểu bang qua các năm

Phân bố doanh thu theo từng tiểu bang



Nhận xét: Có 1 thị trường mà cửa hàng chưa thể tiếp cận được xuyên suốt 4 năm là tiểu bang Alaska ở khu vực miền Tây. Tuy nhiên nhờ doanh thu hơn 300,000 ở tiểu bang California mà miền Tây là khu vực đóng góp nhiều doanh thu nhất trong cửa hàng theo biểu đồ trước đó. California cũng là tiểu bang đóng góp doanh thu nhiều nhất cho cửa hàng xếp sau đó là New York, Texas, Washington.

Tuy miền Tây là khu vực đóng góp nhiều nhất cho cửa hàng nhưng thực chất chỉ đến từ bang California, trong khi đó ở miền Đông hầu như các bang ở đây đều đóng góp vào doanh thu khoảng từ 50,000 – 100,000. Ta thấy được sự cung cấp doanh thu đều đặn ở khu vực này và miền Đông cũng là khu vực thứ 2 đóng góp vào doanh thu cửa hàng chỉ sau miền Tây và đây cũng là lý giải cho việc tăng trưởng ổn định trong doanh thu qua các năm của miền Đông so với các khu vực khác mà ta vừa nhận xét từ biểu đồ trước đó.

4. Kết luận

Như vậy sau khi tìm hiểu bộ dữ liệu thông qua các nhóm biểu đồ, ta đã nắm được một phần về tình hình kinh doanh của cửa hàng xuyên suốt 4 năm từ đầu năm 2014 đến đầu năm 2018.

Như là trong năm 2015, cửa hàng đã trải qua một sự kiện khó khăn điều này thể hiện qua cả lượng hàng bán được, doanh thu, lợi nhuận. Tuy vậy, có lẽ nhờ khả năng của các nhà quản trị tình hình kinh doanh của họ chỉ chậm lại nhưng chưa quá mức thụt lùi và các năm sau đó họ vẫn có thể phát triển một cách mạnh mẽ.

Bên cạnh đó, ta thấy được quy mô của cửa hàng đã trải rộng được hầu như tất cả phần lãnh thổ của nước Mỹ ngoại trừ bang Alaska. Tuy nhiên, chỉ có một vài trong số đó là thị trường chủ chốt mà họ đã thành công chinh phục và mang lại lượng lớn doanh thu và lợi nhuận điển hình là bang California.

Đồng thời ta thấy được tuy trong giai đoạn 2015 có sự chậm lại trong kinh doanh nhưng ở khu vực phía Đông (East) thì tình hình kinh doanh vẫn luôn phát triển một cách ổn định. Và ta đã thấy được nguyên nhân chính là do việc chỉ tập trung một vài thị trường dẫn đến khi có sự kiện đột xuất thì ở các khu vực khác đều chịu ảnh hưởng, ngược lại ở phía Đông do hầu hết các bang đều đóng góp doanh thu ở mức tốt do vậy thị trường này vẫn phát triển ổn định mà không chịu ảnh hưởng trong sự kiện này.

Ngoài ra còn những thông tin khác mà nhóm em đã đề cập ở các biểu đồ trước và cũng còn nhiều thông tin mà nhóm em vẫn chưa thể khai thác hết trong bộ dữ liệu này. Dẫu vậy em nghĩ ta cũng đã có được một cái nhìn tổng quát về bộ dữ liệu. Đồng thời, qua đó còn là kinh nghiệm khi khai thác thông tin trong một bộ dữ liệu bằng cách biểu diễn thành các biểu đồ.

Danh sách thành viên

Mã số sinh viên	Họ và tên	Đóng góp
31211027658	Nguyễn Quang Nhật	100%
31211027664	Nguyễn Nhật Quang	100%
31211027669	Văn Ngọc Như Quỳnh	100%
31211027660	Nguyễn Thị Ngọc Nhi	100%
31211027667	Đào Thị Phương Quỳnh	100%