PHÂN TÍCH DỮ LIỆU BÁN HÀNG CỦA DOANH NGHIỆP

THÀNH VIÊN TRONG NHÓM

NGUYỄN HOÀI LÂM	3120410274
NGUYỄN PHAN HUY LƯỢNG	3120410312

NỘI DUNG

- MÔ TẢ VỀ ĐỀ TÀI
- MÔ TẢ DỮ LIỆU
- TIỀN XỬ LÝ DỮ LIỆU
- PHÂN TÍCH DỮ LIỆU
- KẾT LUẬN

MÔ TẢ VỀ ĐỀ TÀI

MÔ TẢ VỀ ĐỀ TÀI

Đề tài "Phân tích dữ liệu bán hàng của doanh nghiệp" là một chủ đề quan trọng và hấp dẫn trong lĩnh vực quản lý kinh doanh và phân tích dữ liệu. Nó tập trung vào việc sử dụng công nghệ và phương pháp phân tích dữ liệu để hiểu rõ hơn về quá trình bán hàng của một doanh nghiệp và từ đó đưa ra các quyết định chiến lược. Báo cáo này xoay quanh việc phân tích dữ liệu bán hàng của doanh nghiệp.

Mục tiêu của đề tài là cung cấp cho các doanh nghiệp những kiến thức và kỹ năng cần thiết để phân tích dữ liệu bán hàng hiệu quả. Cụ thể, mục tiêu của đề tài bao gồm:

- Hiểu rõ về khái niệm, vai trò và ứng dụng của phân tích dữ liệu bán hàng.
- Nắm được các loại dữ liệu bán hàng và phương pháp phân tích dữ liệu bán hàng.
- Thực hành phân tích dữ liệu bán hàng bằng các công cụ và phần mềm chuyên dụng.

MÔ TẢ DỮ LIỆU

MÔ TẢ VỀ DỮ LIỆU

File dữ liệu chứa thông tin bán hàng của doanh nghiệp Adidas năm 2020-2021

Nguồn : https://www.kaggle.com/datasets/heemalichaudhari/adidas-sales-dataset/

MÔ TẢ VỀ DỮ LIỆU

Cấu trúc dữ liệu

- **Retail** : Đơn vị hoặc tổ chức bán sản phẩm Adidas
- **Retail ID** : Mã định danh duy nhất cho mỗi nhà bán lẻ
- **Invoice Date** : Ngày diễn ra giao dịch bán hàng
- **Region** : Khu vực địa lý nơi nhà bán lẻ hoạt động
- **State** : Bang trong khu vực nơi đặt trụ sở của nhà bán lẻ
- **City** : Thành phố nơi nhà bán lẻ đặt trụ sở
- **Product** : Sản phẩm Adidas đang được bán
- **Price per Unit** : Giá của một đơn vị sản phẩm Adidas
- **Units Sold** : Số lượng đơn vị sản phẩm Adidas được bán trong một giao dịch cụ thể
- **Total Sales** : Tổng doanh thu được tạo ra từ việc bán sản phẩm Adidas trong một giao dịch
- **Operating Profit** : Lợi nhuận mà nhà bán lẻ kiếm được từ việc bán hàng sau khi trừ chi phí hoạt động
- **Operating Margin** : Tỷ lệ phần trăm lợi nhuận hoạt động so với tổng doanh thu
- **Sales Method** : Phương thức hoặc kênh mà qua đó giao dịch bán hàng diễn ra

TIỀN XỬ LÝ DỮ LIỆU

TIỀN XỬ LÝ DỮ LIỆU

Tiến hành Import thư viện pandas để đọc file dữ liệu

```
import pandas as pd
```

Câu lệnh đọc file dữ liệu

```
filename ="Adidas US Sales Datasets.xlsx"
df = pd.read_excel(filename)
df.head()
```

TIỀN XỬ LÝ DỮ LIỆU

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Ur
0	NaN	NaN	Adidas Sales Database	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	NaN	Retailer	Retailer ID	Invoice Date	Region	State	City	Product	Price per Unit	Units Sold	Total Sales	Operating Profit	Operating Margin	
4	NaN	Foot Locker	1185732	2020-01-01 00:00:00	Northeast	New York	New York	Men's Street Footwear	50	1200	600000	300000	0.5	

Nhận thấy dữ liệu còn chứa những khoảng trống không có dữ liệu (dữ liệu chưa được sạch), chúng ta tiến hành làm sạch dữ liệu

```
df = df.drop(df.index[0:3]) #xóa đi 3 dòng đầu tiên
df.drop("Unnamed: 0", axis = 1, inplace = True) #drop cột đầu tiên
df.columns = df.iloc[0] #Thay đổi tên các cột
df = df.drop(df.index[0]) #Bỏ hàng đầu tiên vì nó chứa tên cột
df = df.reset_index(drop=True) #Đặt lại chỉ mục và xóa chỉ mục hiện có
df.duplicated().sum() #Kiểm tra xem tập dữ liệu có giá trị trùng lặp hay không
```

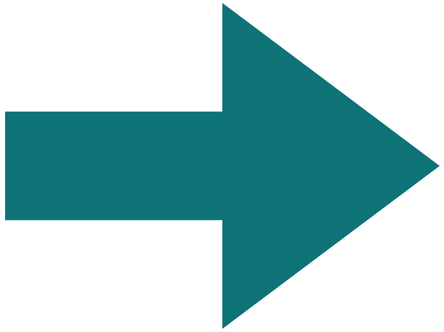
TIỀN XỬ LÝ DỮ LIỆU

3	Retailer	Retailer ID	Invoice Date	Region	State	City	Product	Price per Unit	Units Sold	Total Sales	Operating Profit	Operating Margin	Sales Method
0	Foot Locker	1185732	2020-01-01 00:00:00	Northeast	New York	New York	Men's Street Footwear	50	1200	600000	300000	0.5	In-store
1	Foot Locker	1185732	2020-01-02 00:00:00	Northeast	New York	New York	Men's Athletic Footwear	50	1000	500000	150000	0.3	In-store
2	Foot Locker	1185732	2020-01-03 00:00:00	Northeast	New York	New York	Women's Street Footwear	40	1000	400000	140000	0.35	In-store
3	Foot Locker	1185732	2020-01-04 00:00:00	Northeast	New York	New York	Women's Athletic Footwear	45	850	382500	133875	0.35	In-store
4	Foot Locker	1185732	2020-01-05 00:00:00	Northeast	New York	New York	Men's Apparel	60	900	540000	162000	0.3	In-store

Dữ liệu sau khi đã được làm sạch

TIỀN XỬ LÝ DỮ LIỆU

```
df.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9648 entries, 0 to 9647
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Retailer          9648 non-null   object
 1   Retailer ID       9648 non-null   object
 2   Invoice Date      9648 non-null   object
 3   Region            9648 non-null   object
 4   State             9648 non-null   object
 5   City              9648 non-null   object
 6   Product           9648 non-null   object
 7   Price per Unit    9648 non-null   object
 8   Units Sold        9648 non-null   object
 9   Total Sales       9648 non-null   object
 10  Operating Profit  9648 non-null   object
 11  Operating Margin  9648 non-null   object
 12  Sales Method      9648 non-null   object
dtypes: object(13)
memory usage: 980.0+ KB
```



Thay đổi kiểu dữ liệu cho các cột

```
df.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9648 entries, 0 to 9647
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Retailer          9648 non-null   object
 1   Retailer ID       9648 non-null   object
 2   Invoice Date      9648 non-null   datetime64[ns]
 3   Region            9648 non-null   object
 4   State             9648 non-null   object
 5   City              9648 non-null   object
 6   Product           9648 non-null   object
 7   Price per Unit    9648 non-null   float64
 8   Units Sold        9648 non-null   float64
 9   Total Sales       9648 non-null   float64
 10  Operating Profit  9648 non-null   float64
 11  Operating Margin  9648 non-null   float64
 12  Sales Method      9648 non-null   object
dtypes: datetime64[ns](1), float64(5), object(7)
memory usage: 980.0+ KB
```

PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH DỮ LIỆU

CÂY QUYẾT ĐỊNH

Cây quyết định là một mô hình trong machine learning thuộc họ supervised learning, được sử dụng cho cả vấn đề phân loại (classification) và hồi quy (regression).

Cách Cây Quyết Định Hoạt Động:

1. **Chọn Đặc Trưng:** Chọn đặc trưng (feature) dựa trên tiêu chí nào đó sao cho việc phân loại hoặc dự đoán có tính phân chia tốt nhất.
2. **Chia Dữ Liệu**: Phân tách dữ liệu thành các nhóm con dựa trên giá trị của đặc trưng đã chọn. Mục tiêu là tạo ra các nhóm có tính đồng nhất cao về phân loại hoặc dự đoán.
3. **Lặp Lại Quá Trình**: Lặp lại quá trình trên với từng nhóm con. Tiếp tục chia đến khi một điều kiện dừng được đạt được, chẳng hạn như đạt đến một độ sâu tối đa hoặc không còn có thể phân chia dữ liệu hiệu quả.
4. **Tạo Nút Lá (Leaf Node**): Khi một điều kiện dừng được đạt được, tạo nút lá và gán nhãn (phân loại) hoặc giá trị dự đoán (hồi quy) cho nút lá.
5. **Dự Đoán**: Khi có dữ liệu mới cần dự đoán, mô hình đi qua cây quyết định từ nút gốc đến một nút lá và trả về kết quả của nút lá đó.

PHÂN TÍCH DỮ LIỆU

CÂY QUYẾT ĐỊNH

```
def rmse(targets, predictions):
    return np.sqrt(np.mean(np.square(targets - predictions)))
```

✓ 0.0s

Hàm tính sự chênh lệch giữa giá trị thực tế và giá trị dự đoán trong mô hình hồi quy

```
#Tạo một mô hình cây quyết định
model = tree.DecisionTreeRegressor()

#Chọn đặc trưng và mục tiêu
X = df[['Units Sold','Total Sales','Operating Margin','Price per Unit']] #Đặc trưng
y = df['Operating Profit'] # Mục tiêu

#Sử dụng train_test_split để chia dữ liệu thành bộ huấn luyện và bộ kiểm tra. 80% dữ liệu được sử dụng để huấn luyện (X_train, y_train),
#và 20% để kiểm tra (X_test, y_test). Tham số random_state đảm bảo sự tái tạo của quá trình chia dữ liệu.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

✓ 0.0s

PHÂN TÍCH DỮ LIỆU

```
#Huấn luyện mô hình Cây quyết định
model.fit(X_train, y_train)

#Dự đoán
y_predicted = model.predict(X_test)
y_predicted
```

```
array([2579.2 , 4054.5 , 1298.88, ..., 2716.6 ,  550.8 , 1231.2 ])
```

Units Sold	Total Sales	Operating Margin	Price per Unit	Operating Profit	y_predict
41	7,540	37%	41	2,791	2,716

```
#Đánh giá mô hình bằng R^2 Square
model.score(X_test,y_test)
```

✓ 0.0s

```
0.9986791234654218
```

```
rmse(y_test,y_predicted)
```

✓ 0.0s

```
1951.1985553625793
```

PHÂN TÍCH DỮ LIỆU

HỒI QUY TUYẾN TÍNH

Hồi quy tuyến tính là một trong những phương pháp quan trọng trong machine learning, được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa biến độc lập và biến mục tiêu. Nó được sử dụng chủ yếu cho các vấn đề dự đoán giá trị liên tục, và mục tiêu là tìm ra đường tuyến tính sao cho dự đoán của mô hình gần với giá trị thực tế nhất.

PHÂN TÍCH DỮ LIỆU

HỒI QUY TUYẾN TÍNH

```
def rmse(targets, predictions):
    return np.sqrt(np.mean(np.square(targets - predictions)))
```

✓ 0.0s

```
X = df[['Units Sold','Total Sales','Operating Margin','Price per Unit']] #đặc trưng
y = df['Operating Profit'] # mục tiêu
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25) #chia dữ liệu thành bộ huấn luyện và bộ kiểm tra
#75% dữ liệu được sử dụng để huấn luyện (X_train, y_train), và 25% được sử dụng để kiểm tra (X_test, y_test)
```

✓ 0.0s

PHÂN TÍCH DỮ LIỆU

```
model = LinearRegression()
model.fit(X_train, y_train)
y_predicted = model.predict(X_test)
y_predicted
```

```
array([  4290.8307882 ,  12552.69774818,  69263.50827552, ...,
        21327.45551641,  47792.03658834, 165256.52418341])
```

+ Code + Markdown

```
model.score(X_test,y_test)
```

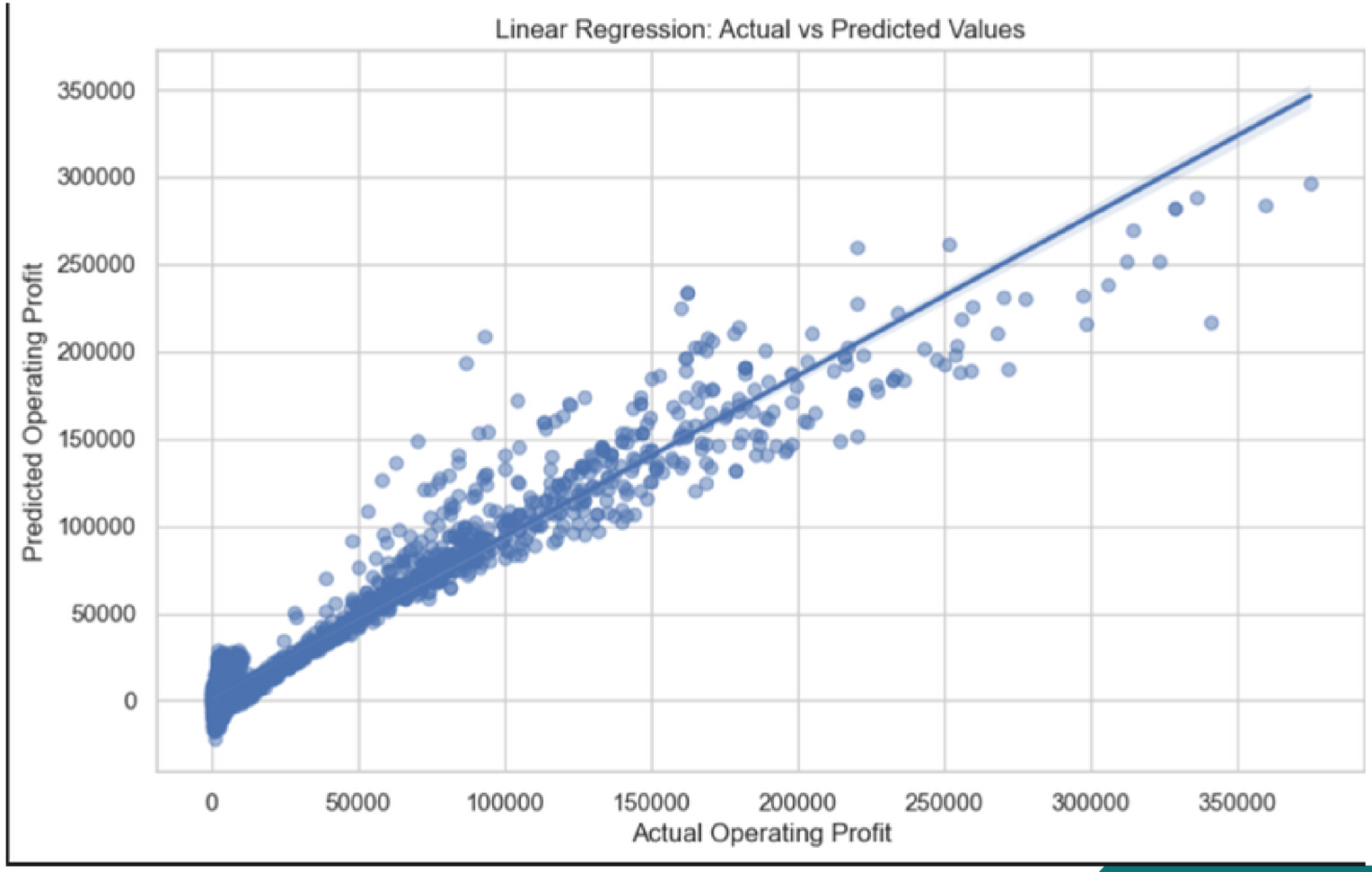
```
0.9395332305632742
```

```
rmse(y_test,y_predicted)
plt.plot()
```

```
13900.815135332881
```

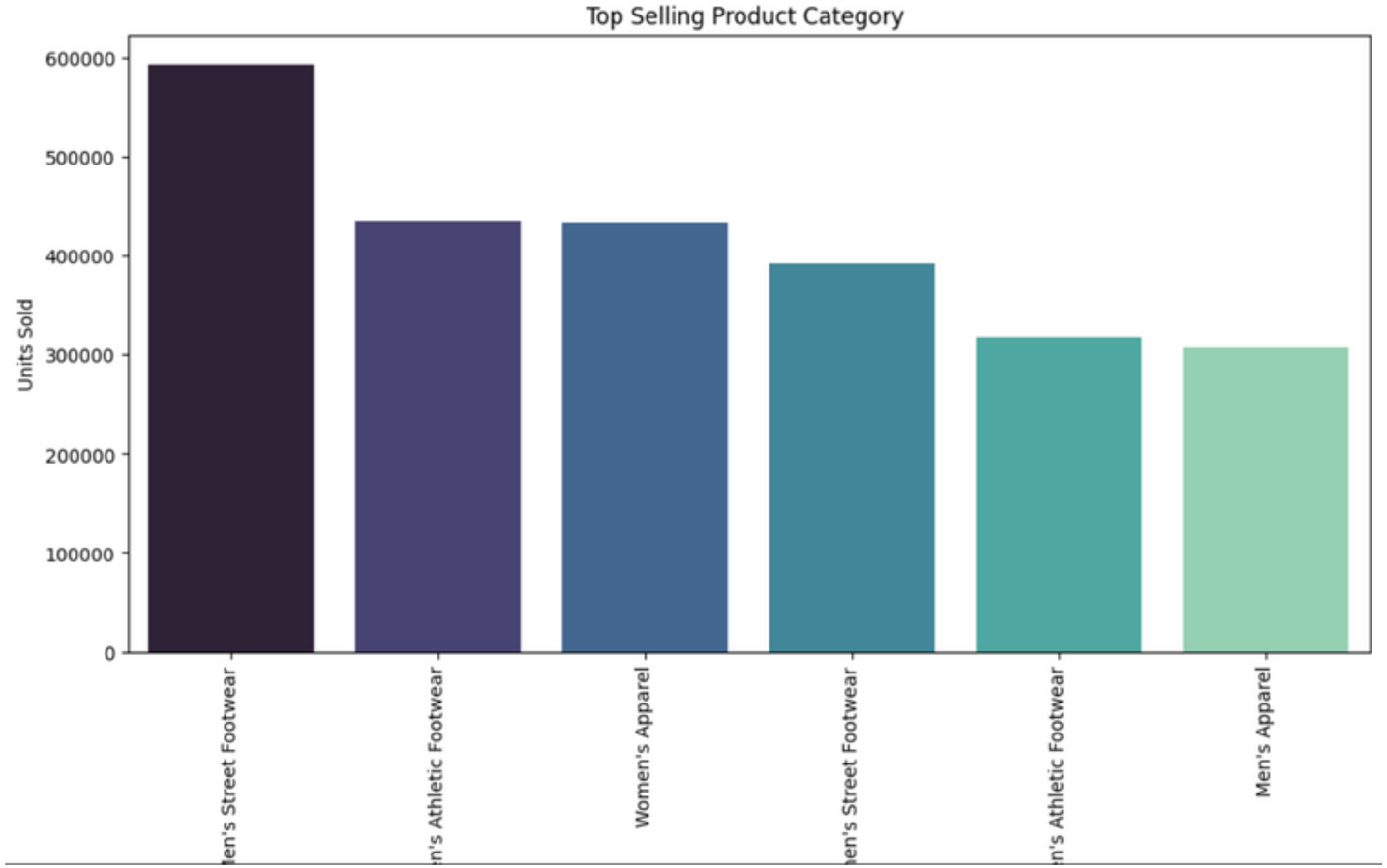
PHÂN TÍCH DỮ LIỆU

HỒI QUY TUYẾN TÍNH



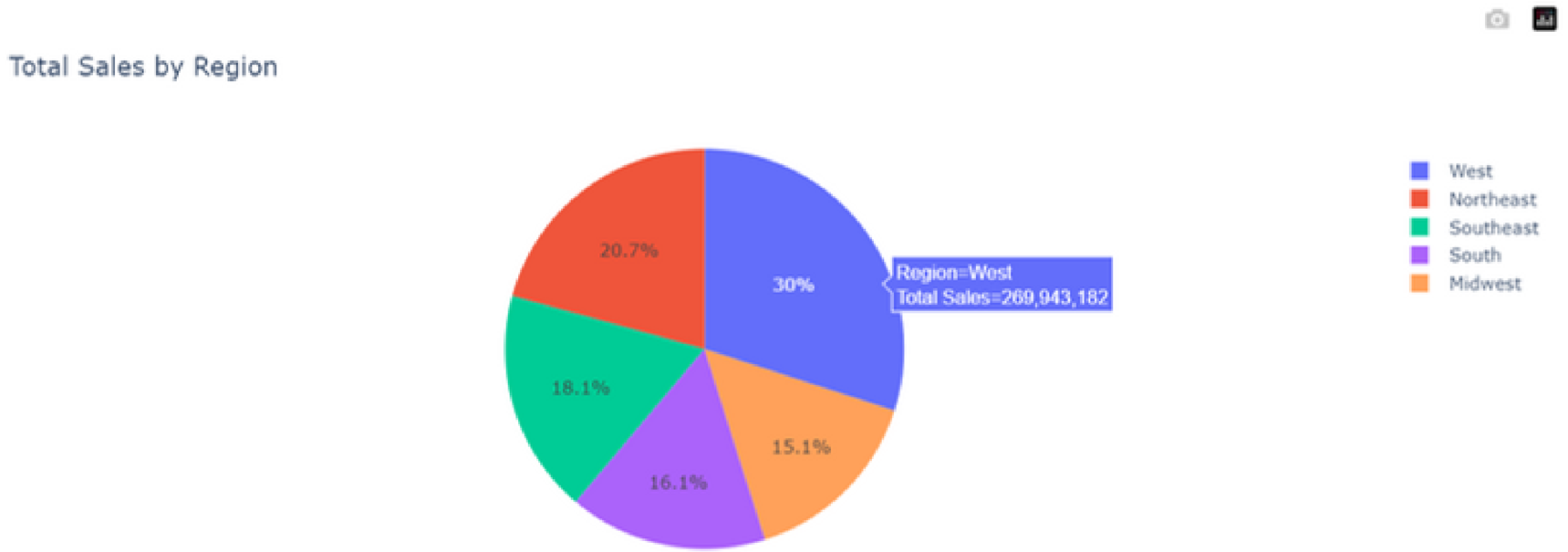
PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH BIỂU ĐỒ



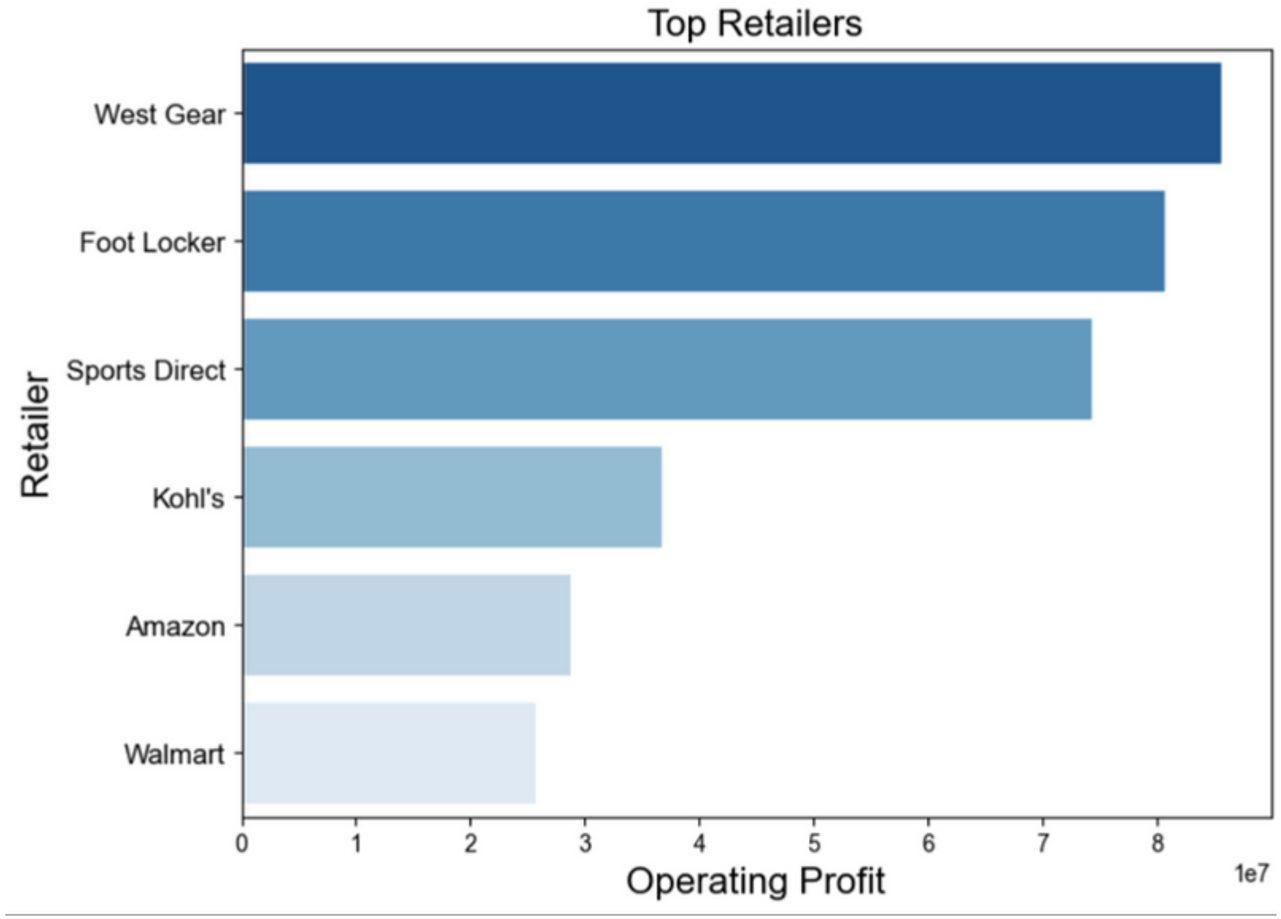
PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH BIỂU ĐỒ



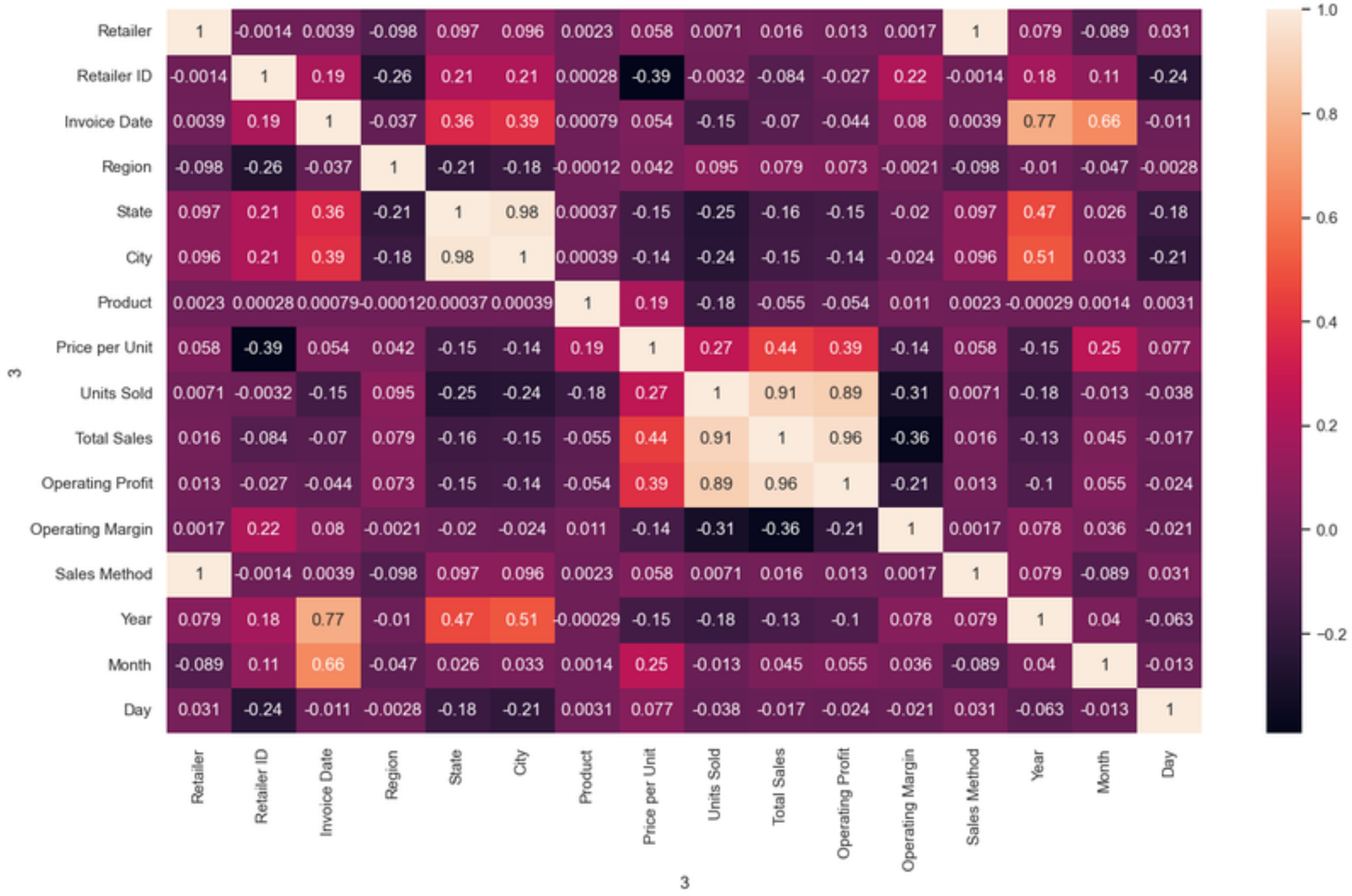
PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH BIỂU ĐỒ



PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH BIỂU ĐỒ



KẾT LUẬN

KẾT LUẬN

Qua bài báo cáo này, nhóm đã hoàn thành mục tiêu đề ra là giải quyết những bài toán liên quan đến phân tích dữ liệu bán hàng của doanh nghiệp, giúp cho doanh nghiệp có thể thuận lợi hơn trong việc kiểm soát tình hình biến động của các sản phẩm được bày bán trên thị trường, từ đó đưa ra những quyết định chiến lược chính xác, hạn chế xác suất thất bại trong dự án. Có thể coi những phương pháp này là cần thiết và vô cùng quan trọng giúp nâng cao hiểu quả phân tích dữ liệu.

THANK YOU