



Phân tích dự đoán doanh thu của siêu thị Walmart



Quy trình làm việc

01. Hiểu về doanh nghiệp và bài toán (Business understanding)
02. Hiểu về dữ liệu (Data understanding)
03. Chuẩn bị dữ liệu (Data preparation)
04. Mô hình hóa (Modeling)
05. Đánh giá (Evaluation)
06. Triển khai (Deployment)



I. Giới thiệu

Trong bài tập lớn này nhóm được Kaggle cung cấp bộ dữ liệu kinh doanh trong quá khứ của siêu thị Walmart và mục tiêu chính sẽ là phân tích khai phá dữ liệu, sử dụng các thuật toán máy học để xây dựng mô hình để dự đoán doanh thu trong tương lai bằng các phương pháp đã được học & tự tìm hiểu

Bài toán cần giải quyết ở đây là dự đoán doanh thu hay nói chung trong lĩnh vực học máy khi mục tiêu là dự đoán một giá trị số liệu dựa trên đặc trưng hoặc biến đầu vào thì bài toán sẽ được gọi là Regression hay còn gọi là hồi quy



II. Hiểu về dữ liệu

Dữ liệu được tải xuống trực tiếp từ trang web Kaggle :
[Walmart Recruiting - Store Sales Forecasting | Kaggle](#).

Dữ liệu được chia thành 4 File CSV bao gồm features.csv, stores.csv, test.csv, train.csv trong đó:

- Features.csv chứa dữ liệu liên quan đến cửa hàng, bộ phận trong những ngày nhất định
- -Stores.csv chứa dữ liệu thông tin cho 45 cửa hàng ngoài ra còn chứa thêm loại cửa hàng và kích cỡ cửa hàng.
- Train.csv chứa dữ liệu huấn luyện từ ngày 2/5/2010 đến ngày 1/11/2012 ngoài ra còn vài thông tin khác:
- -Test.csv chứa những dữ liệu giống với train.csv chỉ thiếu giá trị biến mục tiêu là Weekly_Sales.

1. ĐỌC, QUA 4 FILE DỮ LIỆU & XỬ LÝ ĐỂ PHỤC VỤ CÁC BƯỚC TIẾP THEO:

```
features_data.head()
```

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

```
stores_data.head()
```

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

```
test_data.head()
```

	Store	Dept	Date	IsHoliday
0	1	1	2012-11-02	False
1	1	1	2012-11-09	False
2	1	1	2012-11-16	False
3	1	1	2012-11-23	True
4	1	1	2012-11-30	False

```
train_data.head()
```

	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-02-05	24924.50	False
1	1	1	2010-02-12	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-03-05	21827.90	False

Do các file dữ liệu được tách rời với nhau ra nên nhóm phải thực hiện gộp cái file lại :

```
features_stores_data = features_data.merge(stores_data,how = "inner",on = "Store")
```

```
features_stores_data.head()
```

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday	Type	Size
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False	A	151315
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True	A	151315
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False	A	151315
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False	A	151315
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False	A	151315

```
df_train = train_data.merge(features_stores_data, how='inner', on = ['Store','Date','IsHoliday']).sort_values(by=['Store','Dept',
```

```
tures_stores_data, how='inner', on = ['Store','Date','IsHoliday']).sort_values(by=['Store','Dept','Date']).reset_index(drop=True)
```

Kết quả hai file dữ liệu sau khi đã thực hiện gộp các file và sắp xếp.



df_train.head()

	Store	Dept	Date	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	Markdown1	Markdown2	Markdown3	Markdown4	Markdown5	CPI	Unemploy
0	1	1	2010-02-05	24924.50	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	
1	1	1	2010-02-12	46039.49	True	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	
2	1	1	2010-02-19	41595.55	False	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	
3	1	1	2010-02-26	19403.54	False	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	
4	1	1	2010-03-05	21827.90	False	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	

df_test.head()

	Store	Dept	Date	IsHoliday	Temperature	Fuel_Price	Markdown1	Markdown2	Markdown3	Markdown4	Markdown5	CPI	Unemployment	Type
0	1	1	2012-11-02	False	55.32	3.386	6766.44	5147.70	50.82	3639.90	2737.42	223.462779	6.573	A 15
1	1	1	2012-11-09	False	61.24	3.314	11421.32	3370.89	40.28	4646.79	6154.16	223.481307	6.573	A 15
2	1	1	2012-11-16	False	52.92	3.252	9696.28	292.10	103.78	1133.15	6612.69	223.512911	6.573	A 15
3	1	1	2012-11-23	True	56.23	3.211	883.59	4.17	74910.32	209.91	303.32	223.561947	6.573	A 15
4	1	1	2012-11-30	False	52.34	3.207	2460.03	NaN	3838.35	150.57	6966.34	223.610984	6.573	A 15



2. THỰC HIỆN PHÂN TÍCH VÀ KHAI PHÁ DỮ LIỆU

Kiểm tra các thuộc tính nào là Numeric và các thuộc tính nào là Object

```
for label, content in df_train.items():  
    if not pd.api.types.is_numeric_dtype(content):  
        print(label)
```

Date
Type

```
for label, content in df_train.items():  
    if pd.api.types.is_numeric_dtype(content):  
        print(label)
```

Store
Dept
Weekly_Sales
IsHoliday
Temperature
Fuel_Price
Markdown1
Markdown2
Markdown3
Markdown4
Markdown5
CPI
Unemployment
Size

Kiểm tra xem trong tập dữ liệu có trường hợp bị Null không



```
df_train.isnull().sum()
```

```
Store      0
Dept       0
Date       0
Weekly_Sales  0
IsHoliday  0
Temperature 0
Fuel_Price 0
Markdown1  270889
Markdown2  310322
Markdown3  284479
Markdown4  286603
Markdown5  270138
CPI        0
Unemployment 0
Type       0
Size       0
dtype: int64
```

Kiểm tra các chỉ số của các thuộc tính trong bộ dữ liệu:

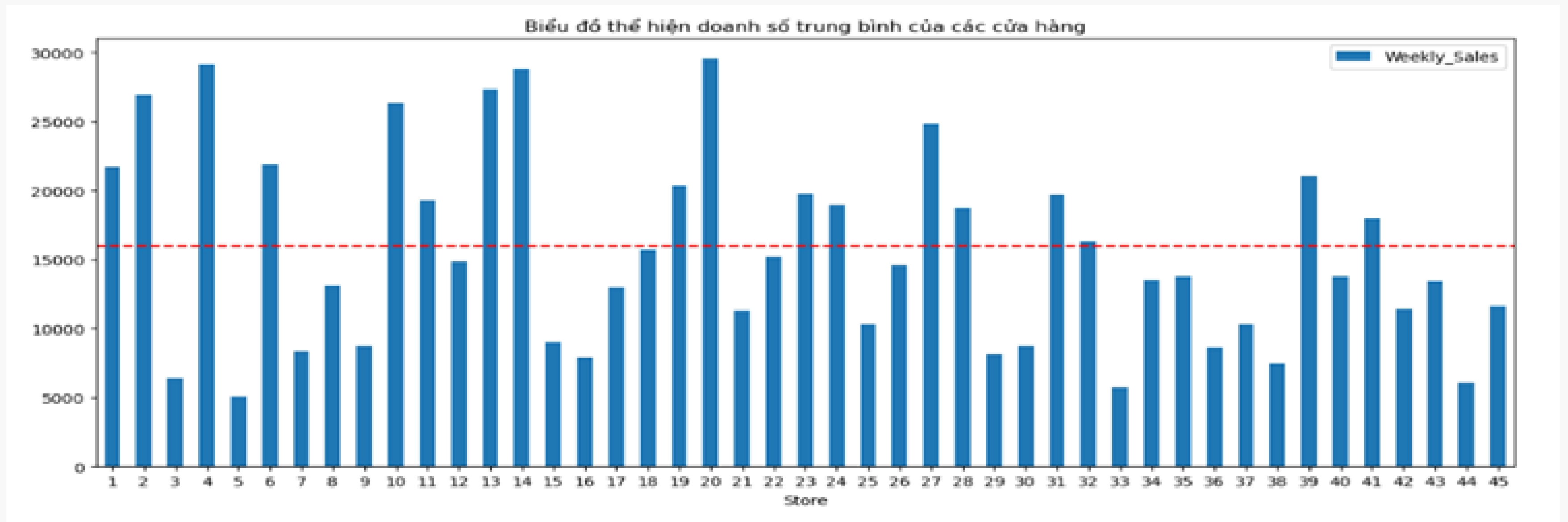
```
df_train.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Store	421570.0	22.200546	12.785297	1.000	11.000000	22.000000	33.000000	45.000000
Dept	421570.0	44.260317	30.492054	1.000	18.000000	37.000000	74.000000	99.000000
Weekly_Sales	421570.0	15981.258123	22711.183519	-4988.940	2079.650000	7612.03000	20205.852500	693099.360000
Temperature	421570.0	60.090059	18.447931	-2.060	46.680000	62.090000	74.280000	100.140000
Fuel_Price	421570.0	3.361027	0.458515	2.472	2.933000	3.45200	3.738000	4.468000
Markdown1	150681.0	7246.420196	8291.221345	0.270	2240.270000	5347.45000	9210.900000	88646.760000
Markdown2	111248.0	3334.628621	9475.357325	-265.760	41.600000	192.00000	1926.940000	104519.540000
Markdown3	137091.0	1439.421384	9623.078290	-29.100	5.080000	24.60000	103.990000	141630.610000
Markdown4	134967.0	3383.168256	6292.384031	0.220	504.220000	1481.31000	3595.040000	67474.850000
Markdown5	151432.0	4628.975079	5962.887455	135.160	1878.440000	3359.45000	5563.800000	108519.280000
CPI	421570.0	171.201947	39.159276	126.064	132.022667	182.31878	212.416993	227.232807
Unemployment	421570.0	7.960289	1.863296	3.879	6.891000	7.86600	8.572000	14.313000
Size	421570.0	136727.915739	60980.583328	34875.000	93638.000000	140167.00000	202505.000000	219622.000000



2.1 Tìm hiểu về thuộc tính Store.

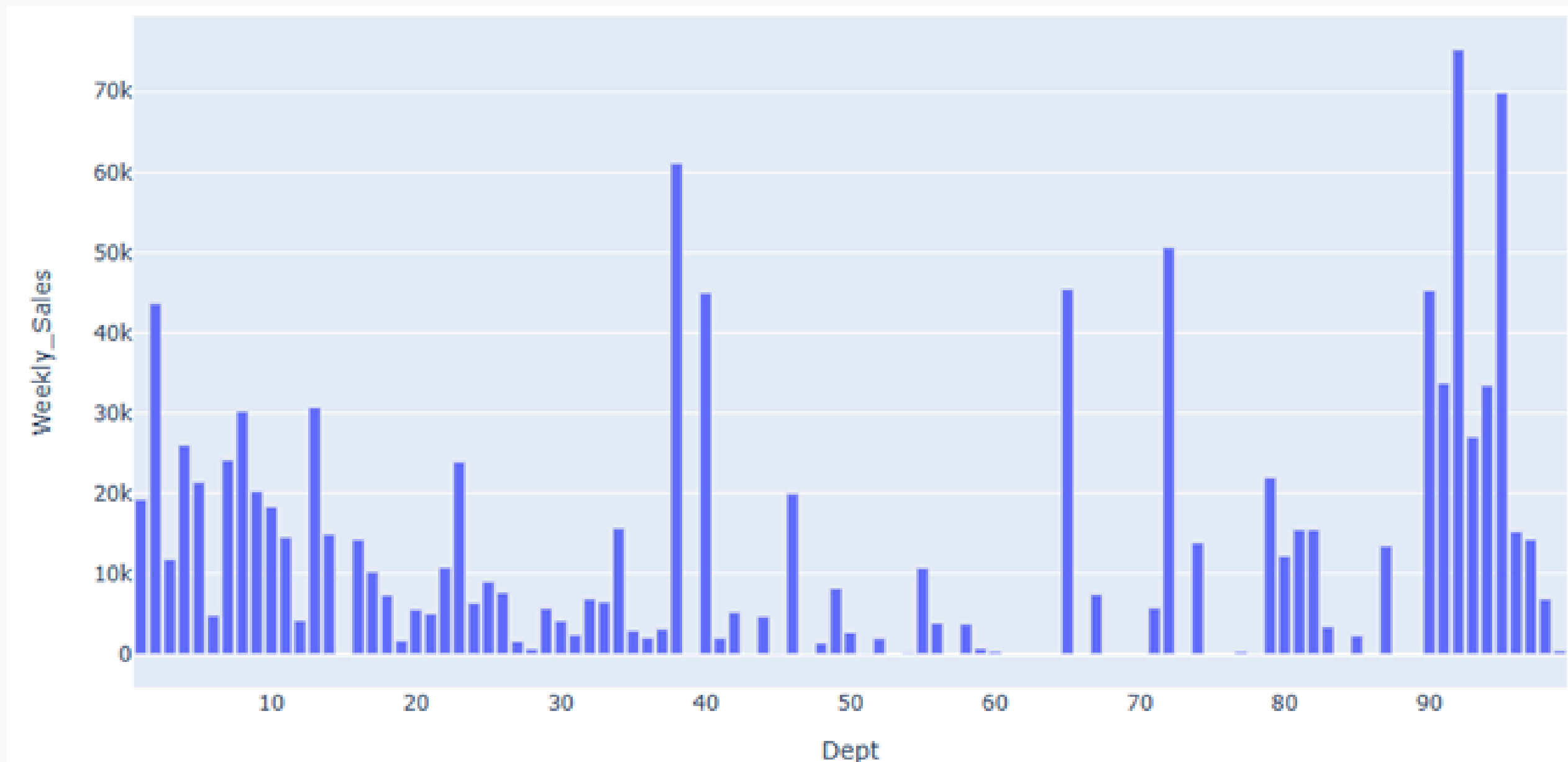
-Có tổng cộng 45 cửa hàng xuất hiện trong bộ dữ liệu.





2.2 Tìm hiểu về thuộc tính Dept

Có 81 phòng ban khác nhau trong bộ dữ liệu.

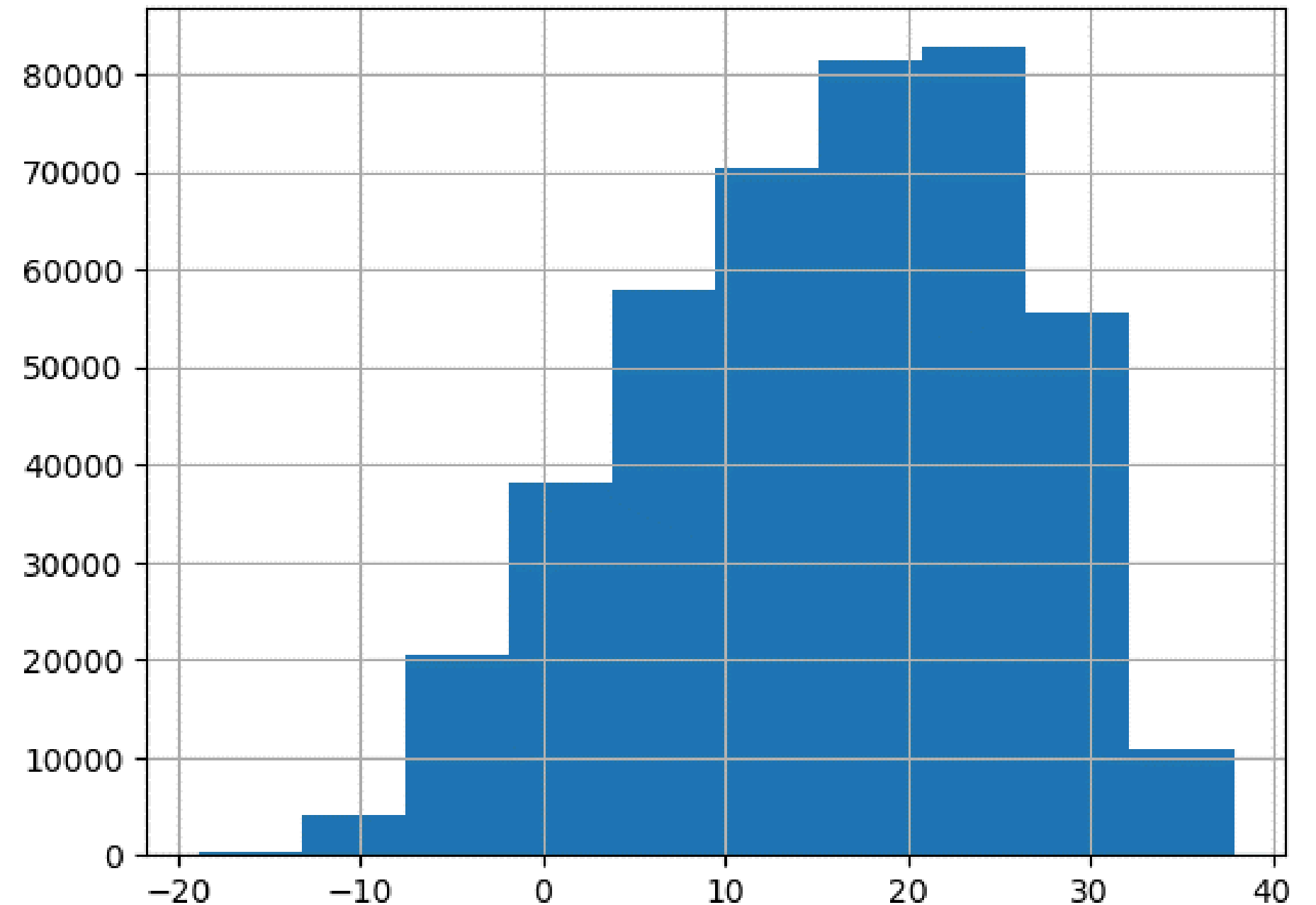


- Doanh thu giữa các phòng ban không đồng đều nhau có những phòng ban tạo ra rất nhiều doanh thu nhưng cũng có các phòng ban tạo ra rất ít doanh thu.

2.3 Tìm hiểu về thuộc tính Temperature

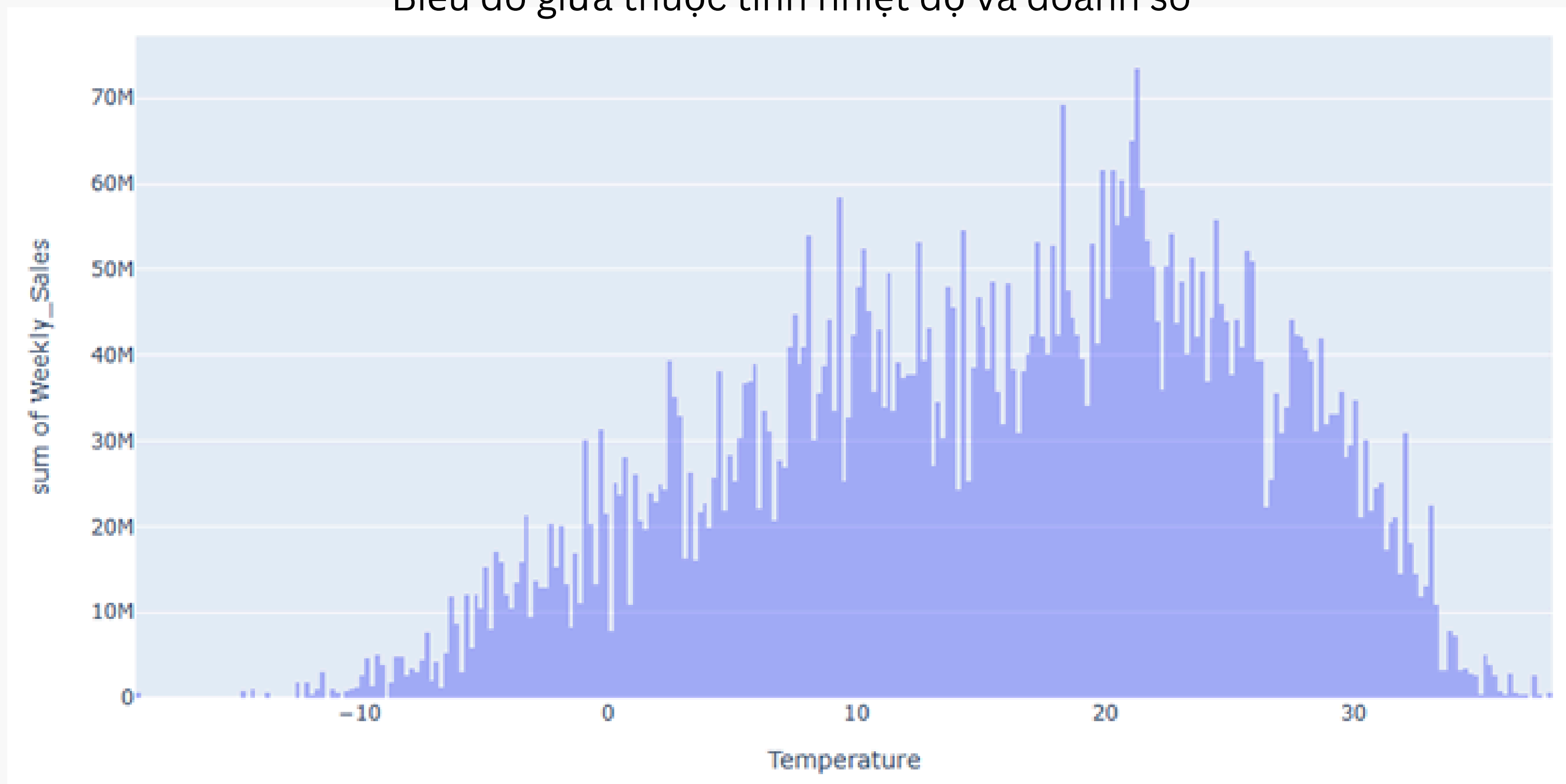
Biểu đồ Histogram về mức độ phân tán của dữ liệu nhiệt độ

- Do nhiệt độ được đo theo độ F nên nhóm sẽ đổi về độ C,
 - Kiểm tra được nhiệt độ trung bình vào khoảng 15 độ
-
- Nhiệt độ chủ yếu dao động ở mức 10 đến dưới 30 độ C
 - Ngoài ra có xuất hiện nhiệt độ dưới 0 độ và ngoài 30.





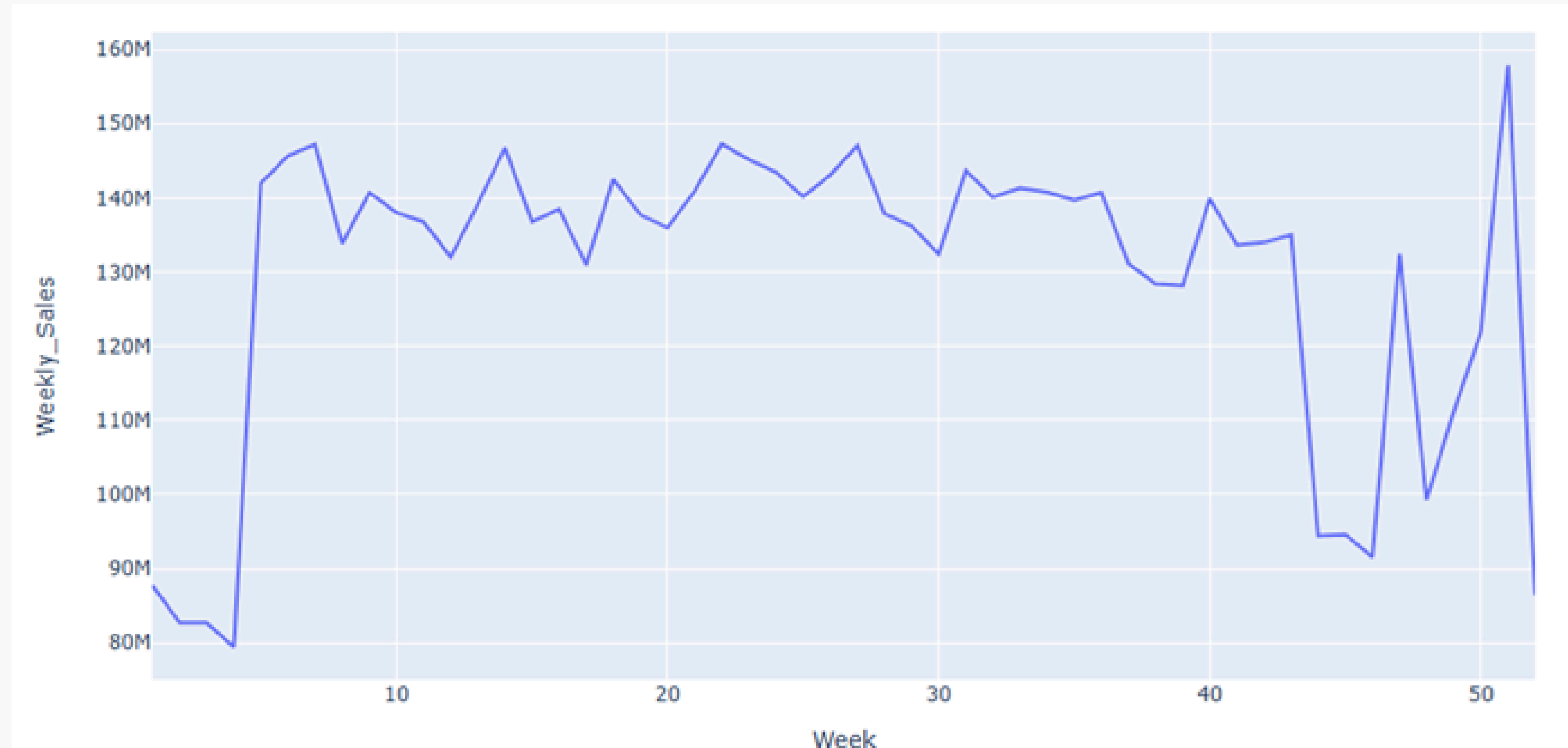
Biểu đồ giữa thuộc tính nhiệt độ và doanh số



Ta thấy những ngày có nhiệt độ ấm áp từ khoảng trên 10 độ và dưới 30 độ người dân sẽ đi mua sắm nhiều hơn từ đó doanh thu cũng tăng

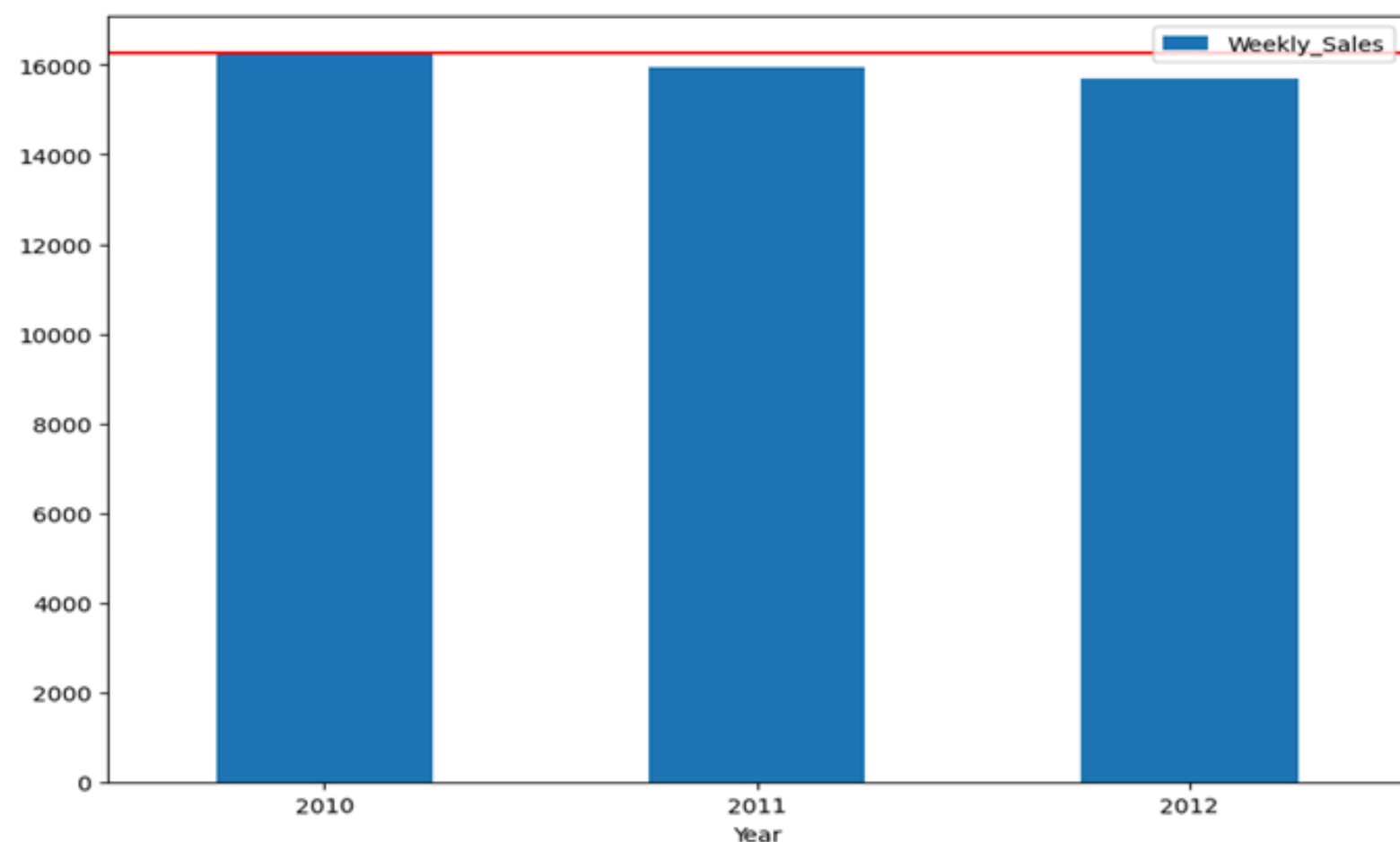
2.4 Tìm hiểu về thuộc tính Date

- Do thuộc tính Date là kiểu dữ liệu object chưa đúng với kiểu dữ liệu ngày tháng nên nhóm phải thực hiện đổi kiểu dữ liệu.
- Ngoài ra nhóm còn tạo ra thêm các cột chứa thêm các thuộc tính ngày, tháng, năm, tuần để phục vụ cho mục đích phân tích và làm đa dạng thêm dữ liệu.



- Từ tuần 44 đến 46 và tuần 48, người dân có thể giảm chi tiêu để tích lũy tiền cho các sự kiện mua sắm lớn hay các chương trình khuyến mãi mà cửa hàng tổ chức vào cuối năm
- Vào cuối năm, đặc biệt là tuần 51 và 52 trong dịp lễ giáng sinh, người dân thường chi tiêu nhiều hơn
- Trong các tuần đầu năm mới như tuần 1, 2, 3 và 4, người dân có thể giảm chi tiêu sau những chi tiêu lớn

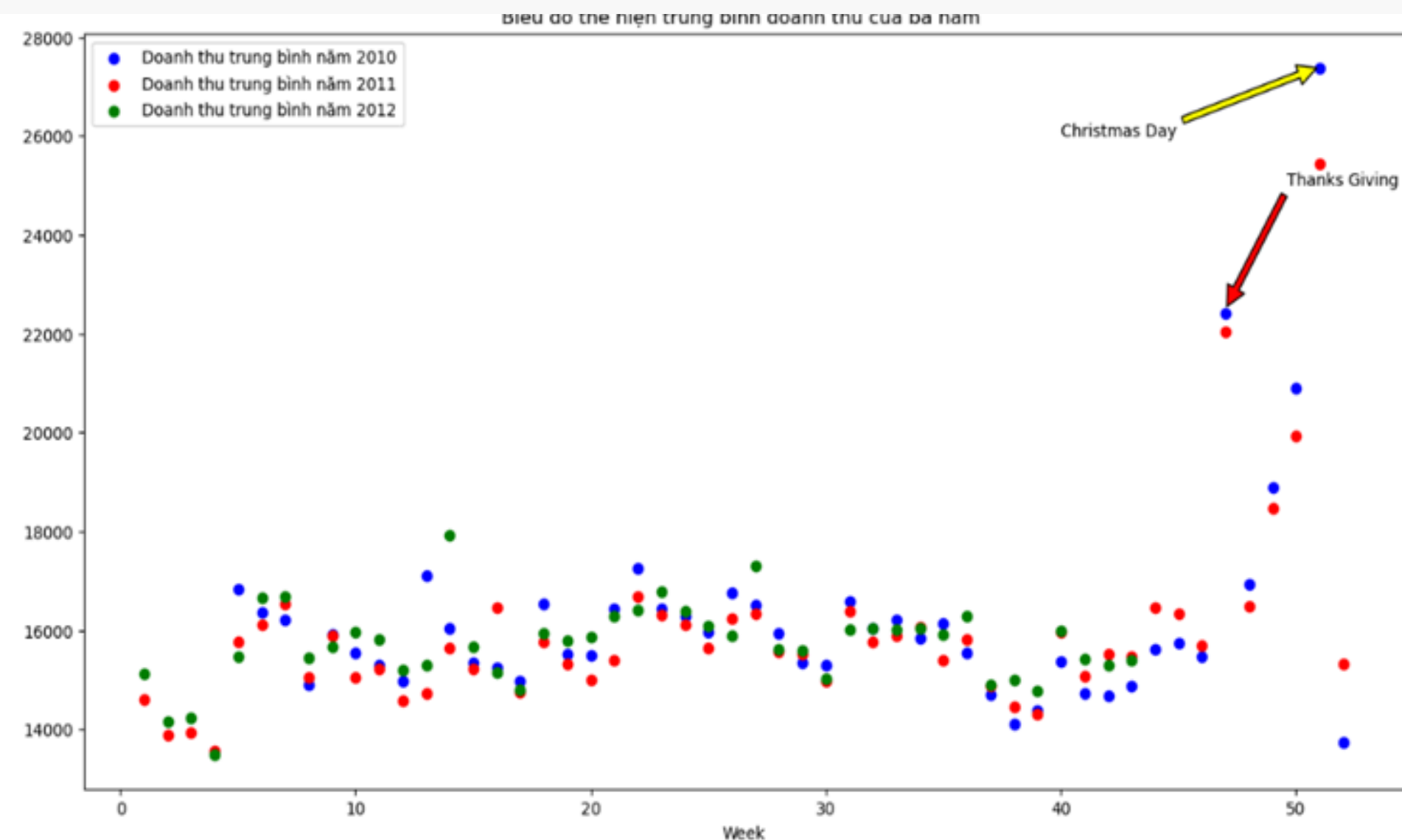
Biểu đồ doanh thu trung bình các năm



- Trong 3 năm trong bộ dữ liệu thì năm 2010 là năm có doanh thu cao nhất trong 3 năm.



Biểu đồ thể hiện doanh thu trung bình tuần của 3 năm



2.5 Tìm hiểu về thuộc tính IsHoliday

Kaggle đã cung cấp thời gian các ngày lễ lớn trong năm có xuất hiện trong bộ dữ liệu. Qua những thông tin trên nhóm đã thực hiện xác định các ngày lễ dựa trên thời gian đã được cung cấp.

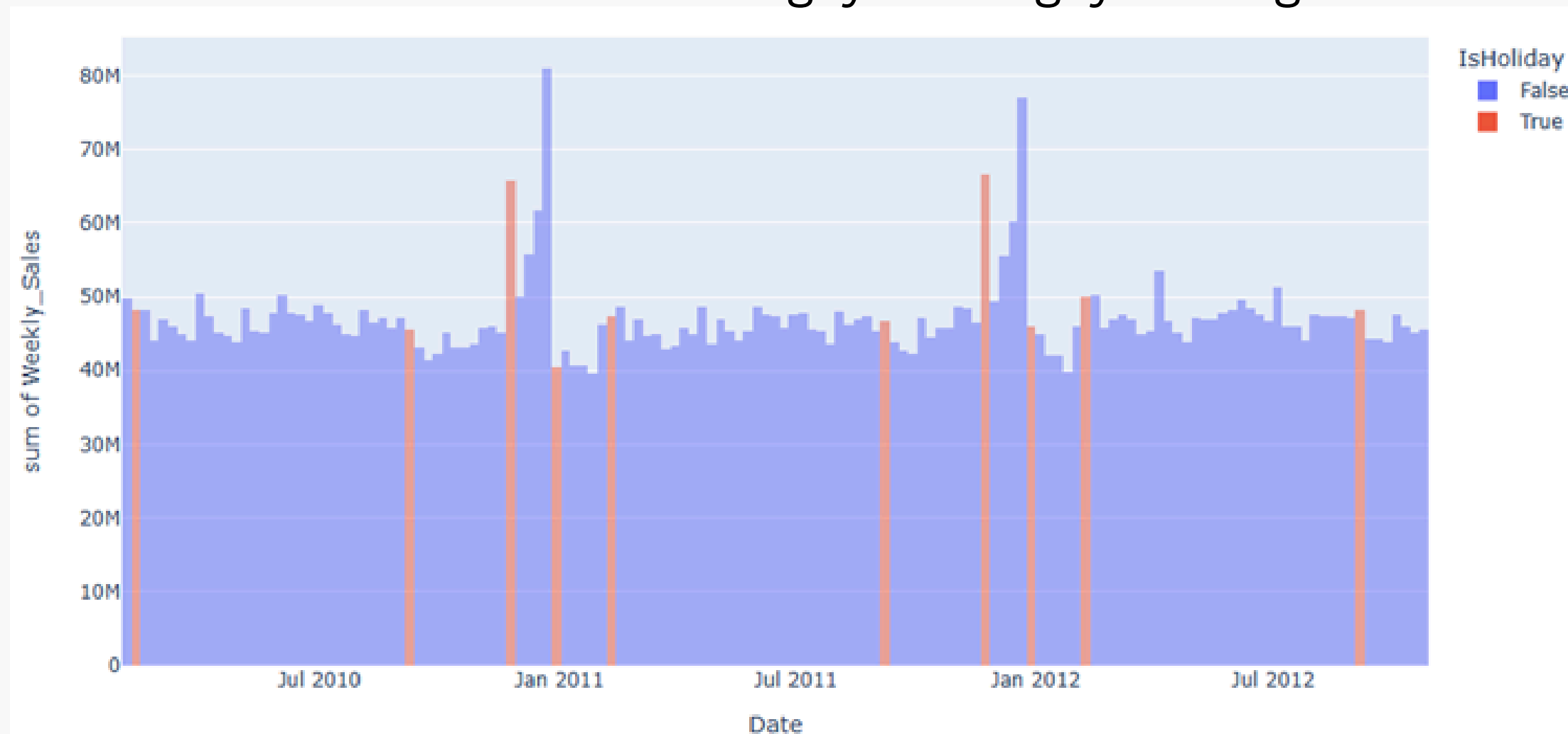
Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13

Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13

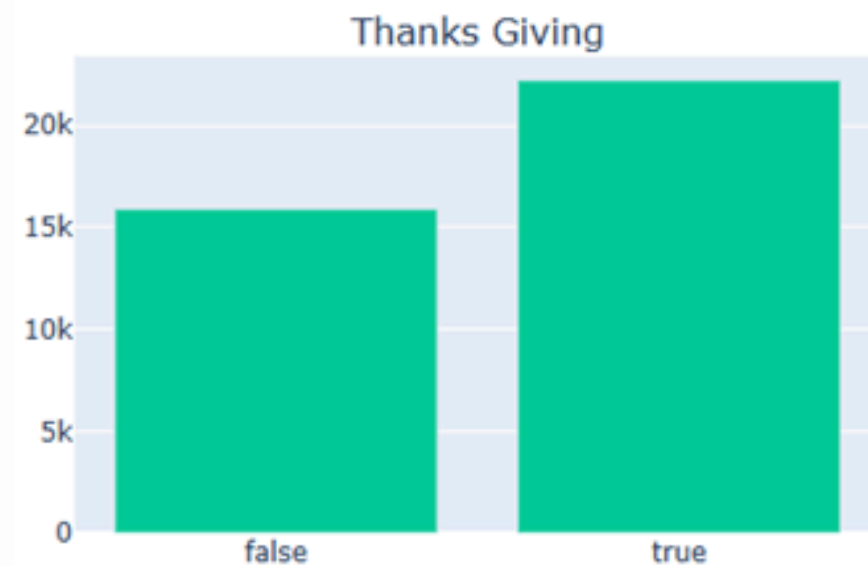
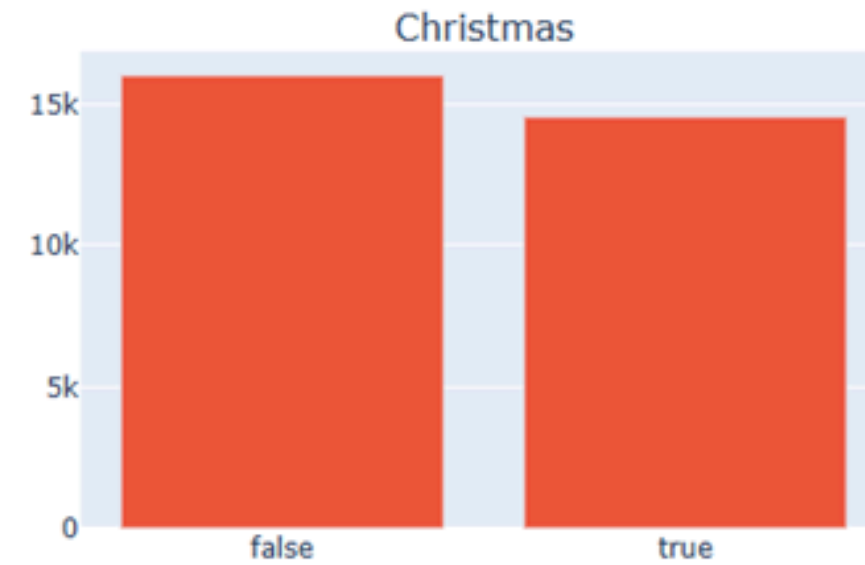
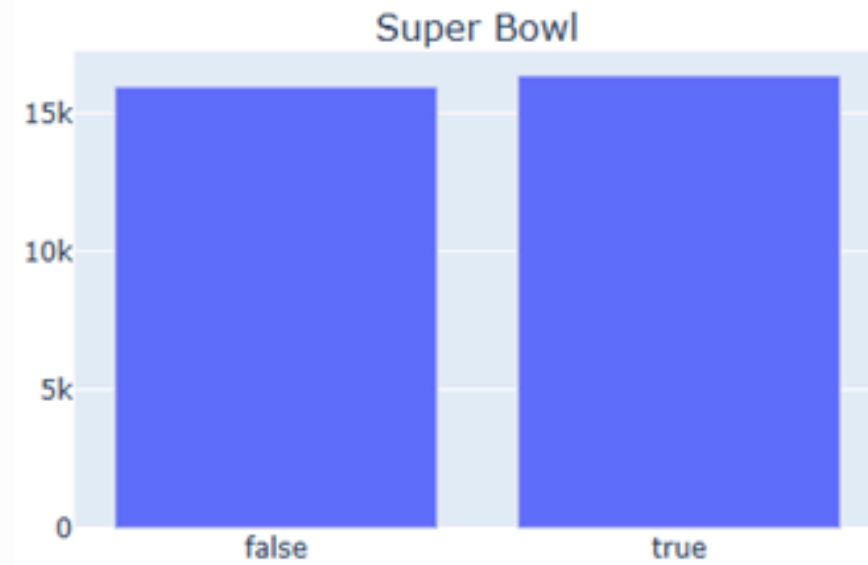
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13

Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Biểu đồ doanh thu vào ngày lễ và ngày thường



Doanh số bán trong 4 ngày lễ so với ngày bình thường

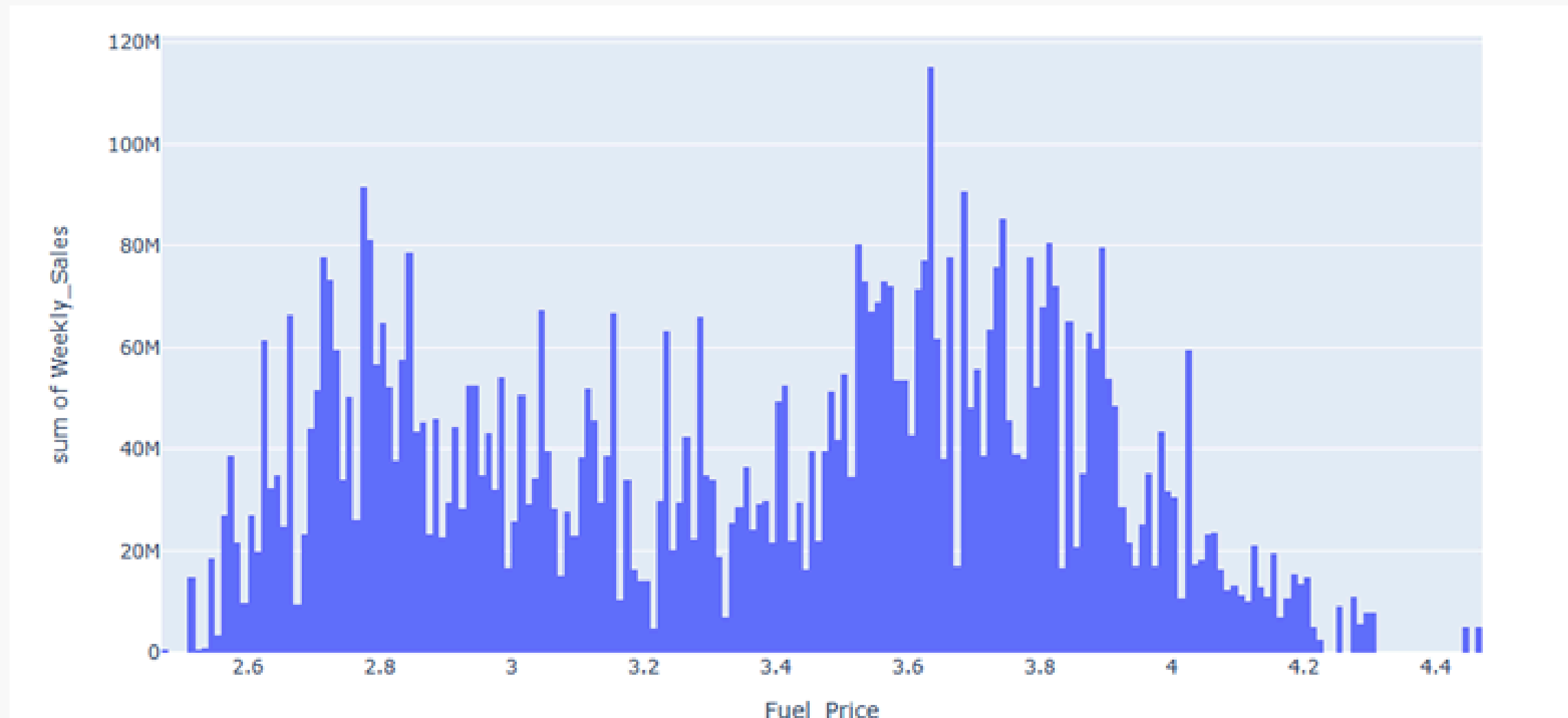


Sau khi quan sát hai biểu đồ so sánh doanh thu giữa ngày thường và ngày lễ trên thì nhóm thấy được có trường hợp tổng doanh thu của những ngày lễ thấp hơn ngày thường có thể do:

- Các cửa hàng thường có các chương trình khuyến mãi lớn vào các ngày lễ để thu hút khách hàng, nhưng nếu tổng lại vẫn thấp hơn so với tổng của nhiều ngày thường.
- Hành vi khách hàng : Có thể do khách hàng có xu hướng mua sắm nhiều hơn vào các ngày thường do các yếu tố khác như sự thoải mái hơn, ít đông đúc hơn
- Các ngày lễ có thể được coi là những dịp để dành thời gian với gia đình và bạn bè hơn là để mua sắm.
- Có sự sai sót trong việc cung cấp thông tin những ngày lễ của kaggle:

2.6 Tìm hiểu về thuộc tính fuel price.

Biểu đồ giá nhiên liệu và doanh thu nhóm



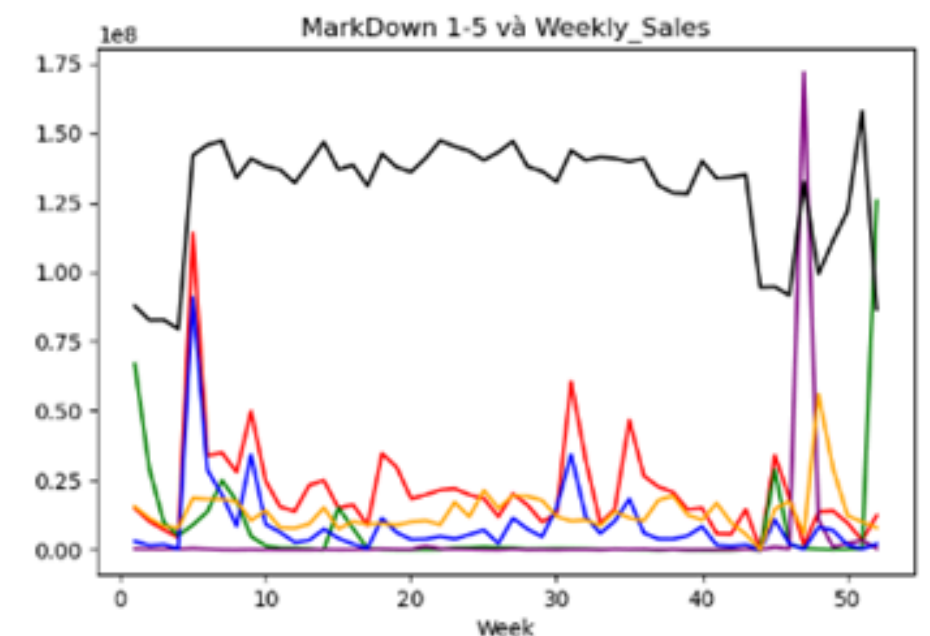
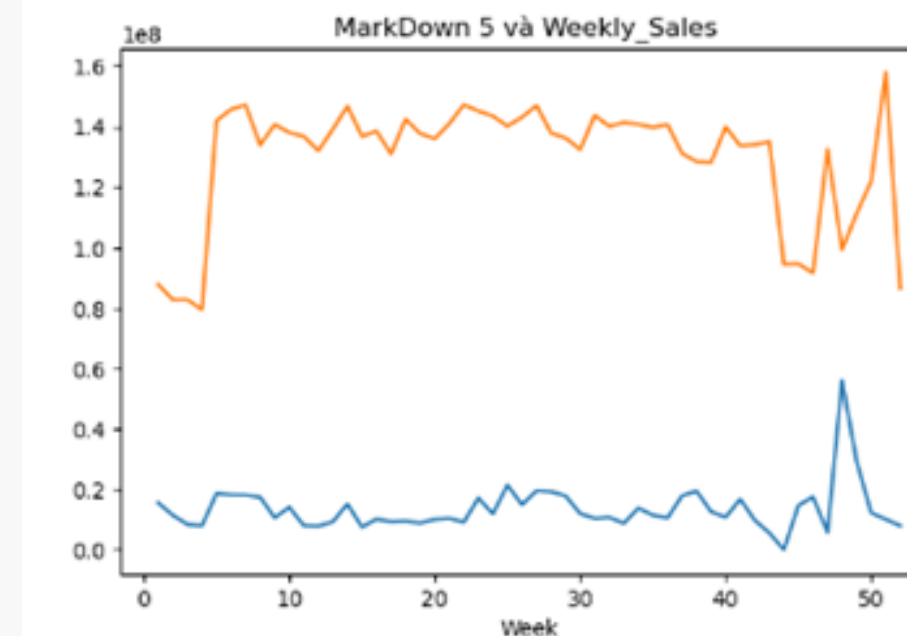
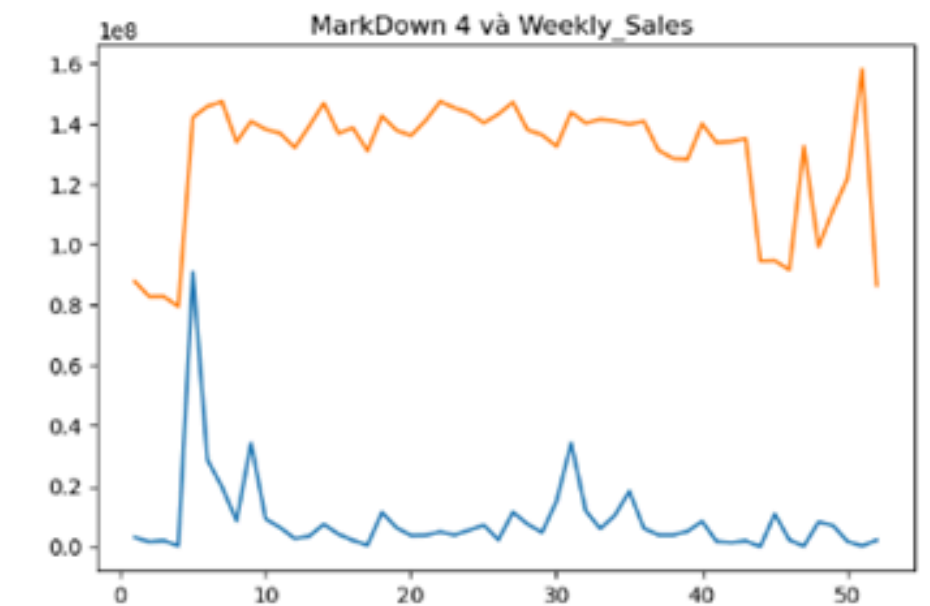
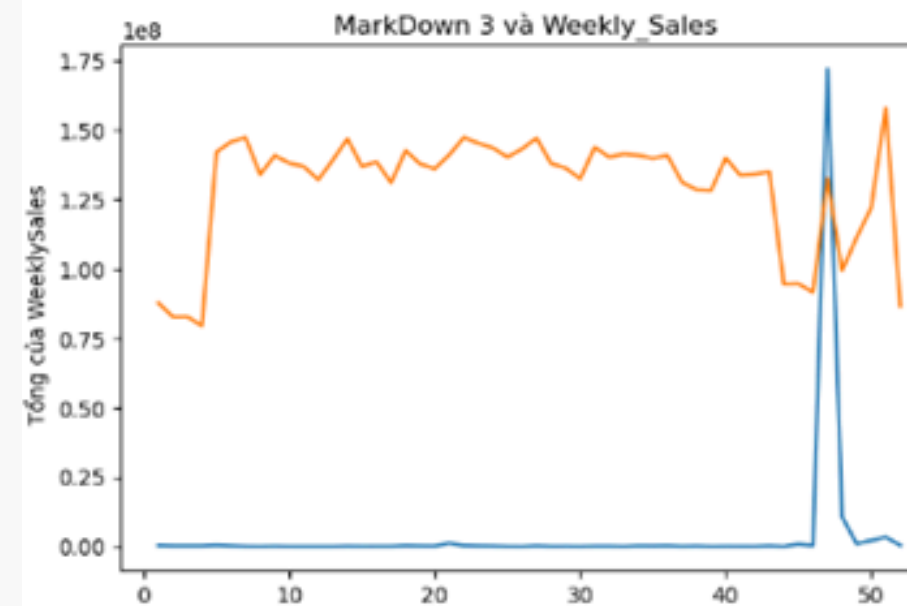
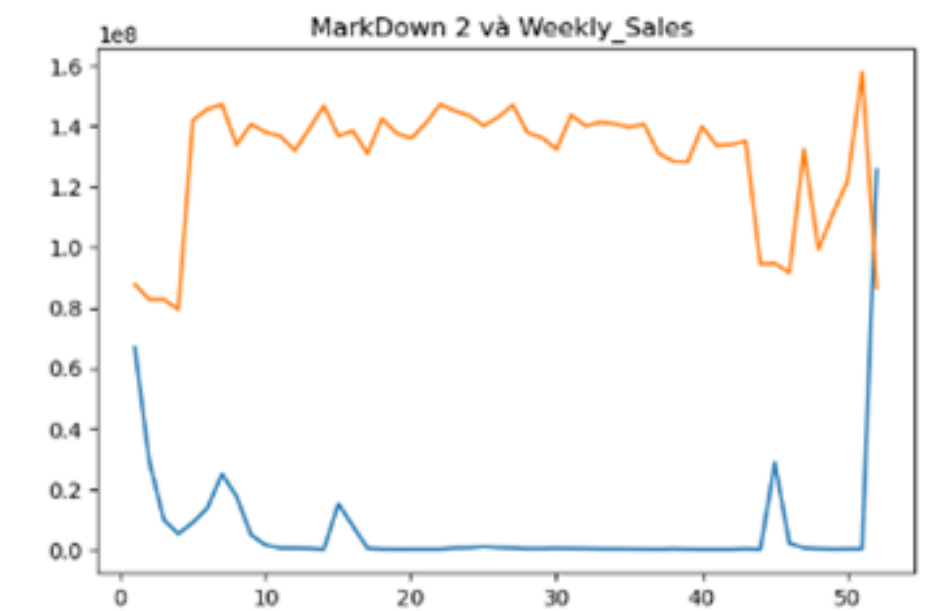
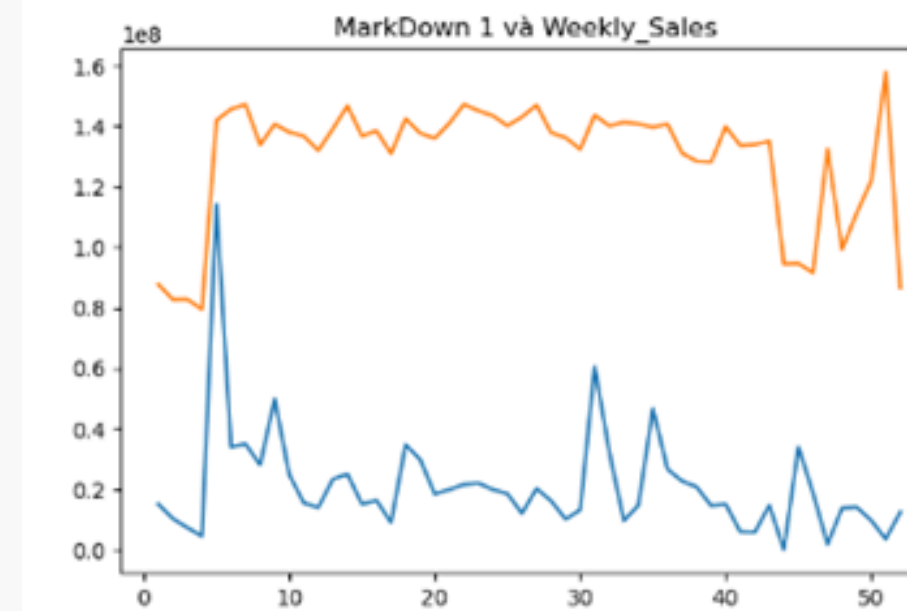
- Việc doanh thu tăng khi giá nhiên liệu giảm có thể do : khi giá nhiên liệu rẻ người dân sẽ đi lại và mua sắm nhiều từ đó làm tăng doanh thu
- Tuy nhiên giá nhiên liệu "thấp" vẫn có trường hợp giảm doanh thu những vẫn có trường hợp tăng doanh thu như ở khoảng 2.3-2.6 và 3-3.5 có thể do mức giá này kéo dài trong thời gian ngắn nên chưa đạt doanh thu cao hoặc cũng có thể do các chiến lược cạnh tranh của các cửa hàng khác
- Ở mức giá nhiên liệu "trung bình" việc mua bán vẫn diễn ra bình thường doanh thu có tăng cũng có giảm không có gì đáng chú ý
- Ở mức giá nhiên liệu "cao" thì doanh thu đã bị giảm do nhiên liệu đắt nên chi phí vận chuyển + sản xuất hàng hóa tăng cao từ đó kéo theo giá sản phẩm cũng tăng cao, ngoài ra khi giá nhiên liệu tăng người dân cũng ít đi lại để tiết kiệm chi phí nên doanh thu bị giảm

2.7 Tìm hiểu về thuộc tính Markdown 1 – 5

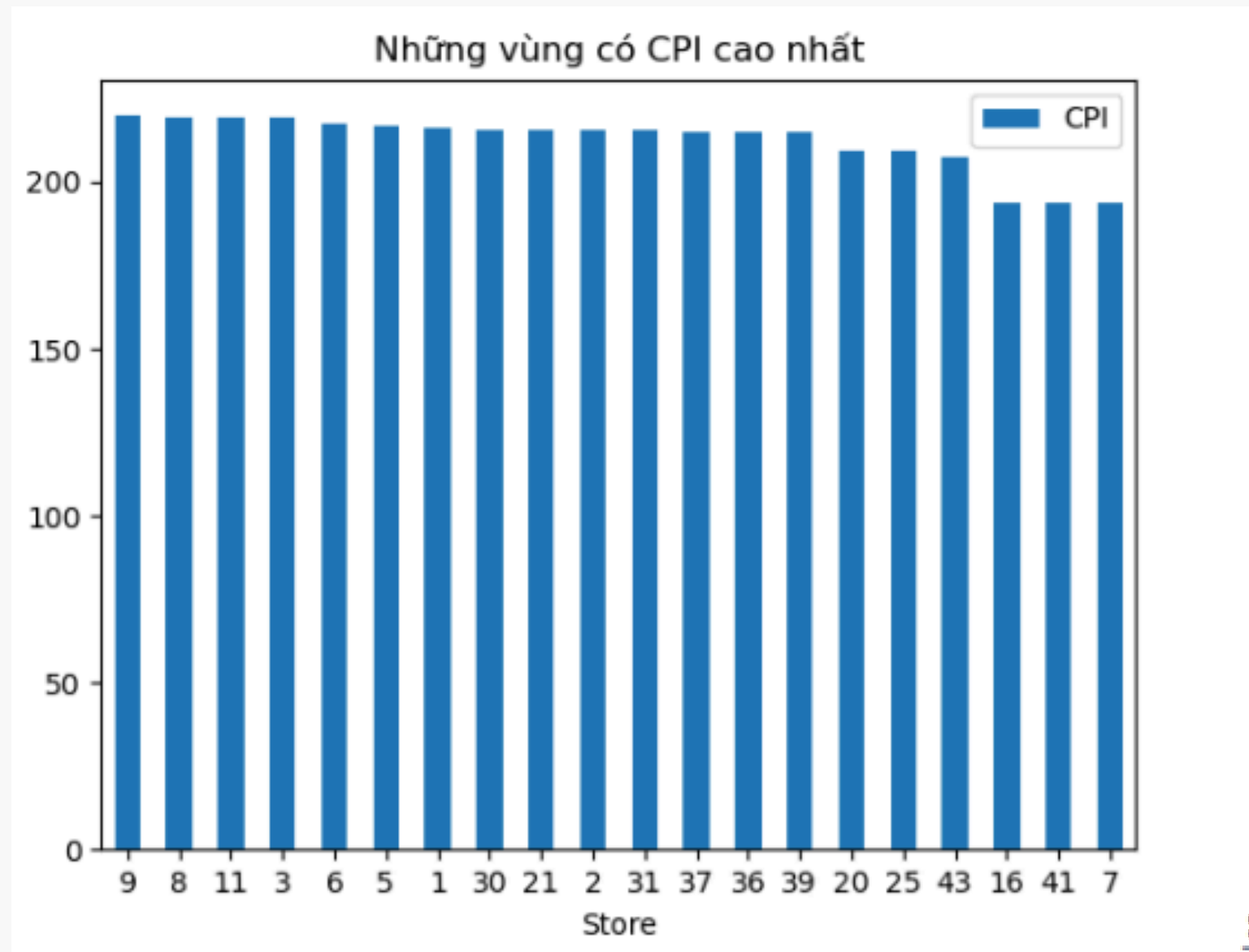
Dữ liệu ẩn về các chiến lược khuyến mãi mà Walmart đang triển khai. Dữ liệu về khuyến mãi chỉ có sẵn sau tháng 11 năm 2011 và không luôn luôn có sẵn cho tất cả các cửa hàng



Theo cá nhân nhóm thực hiện quan sát thấy mỗi lần chỉ số markdown tăng có thể là walmart đang thực hiện một chương trình khuyến mãi từ đó doanh thu sau đó cũng tăng theo.

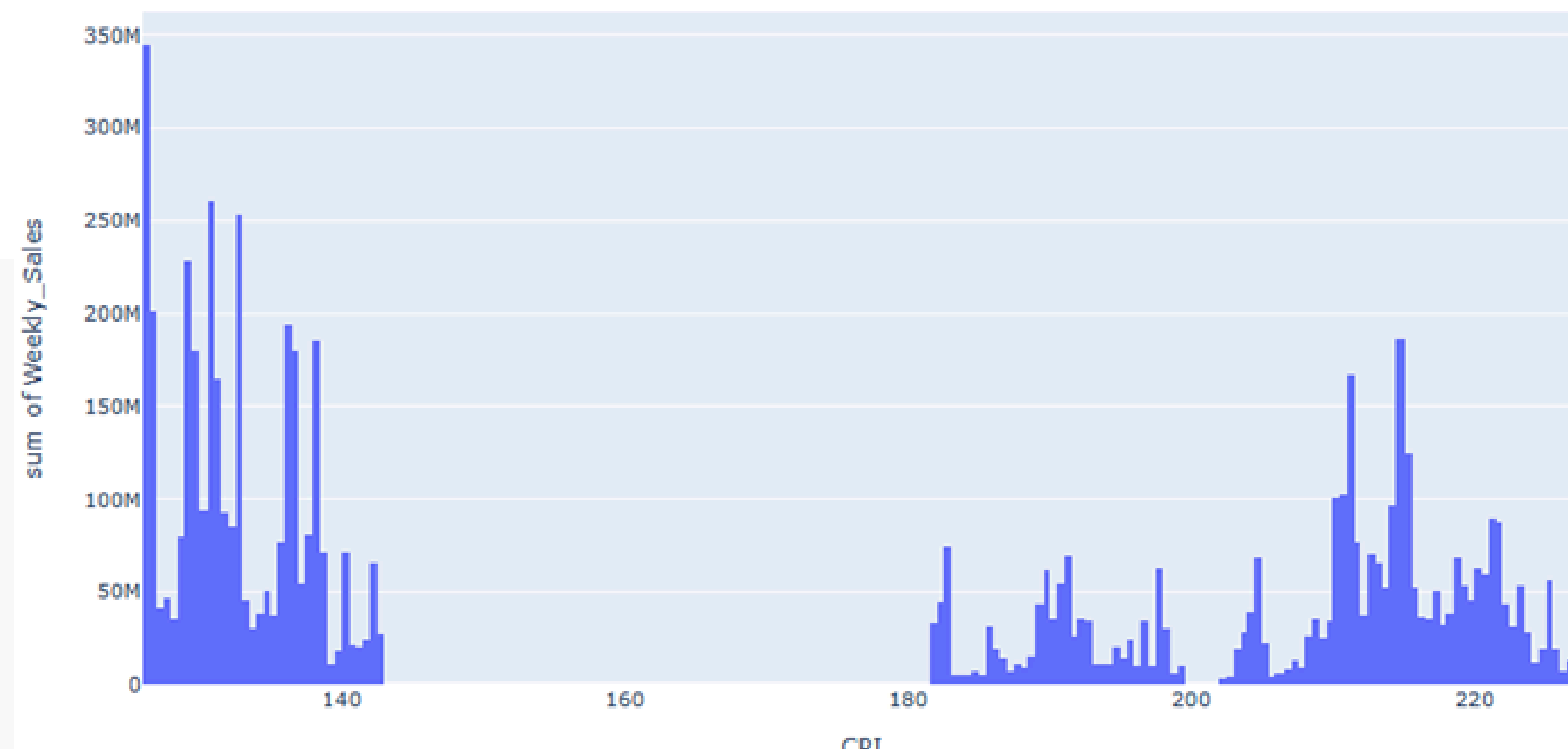


2.8 Tìm hiểu về thuộc tính CPI (chỉ số giá tiêu dùng)

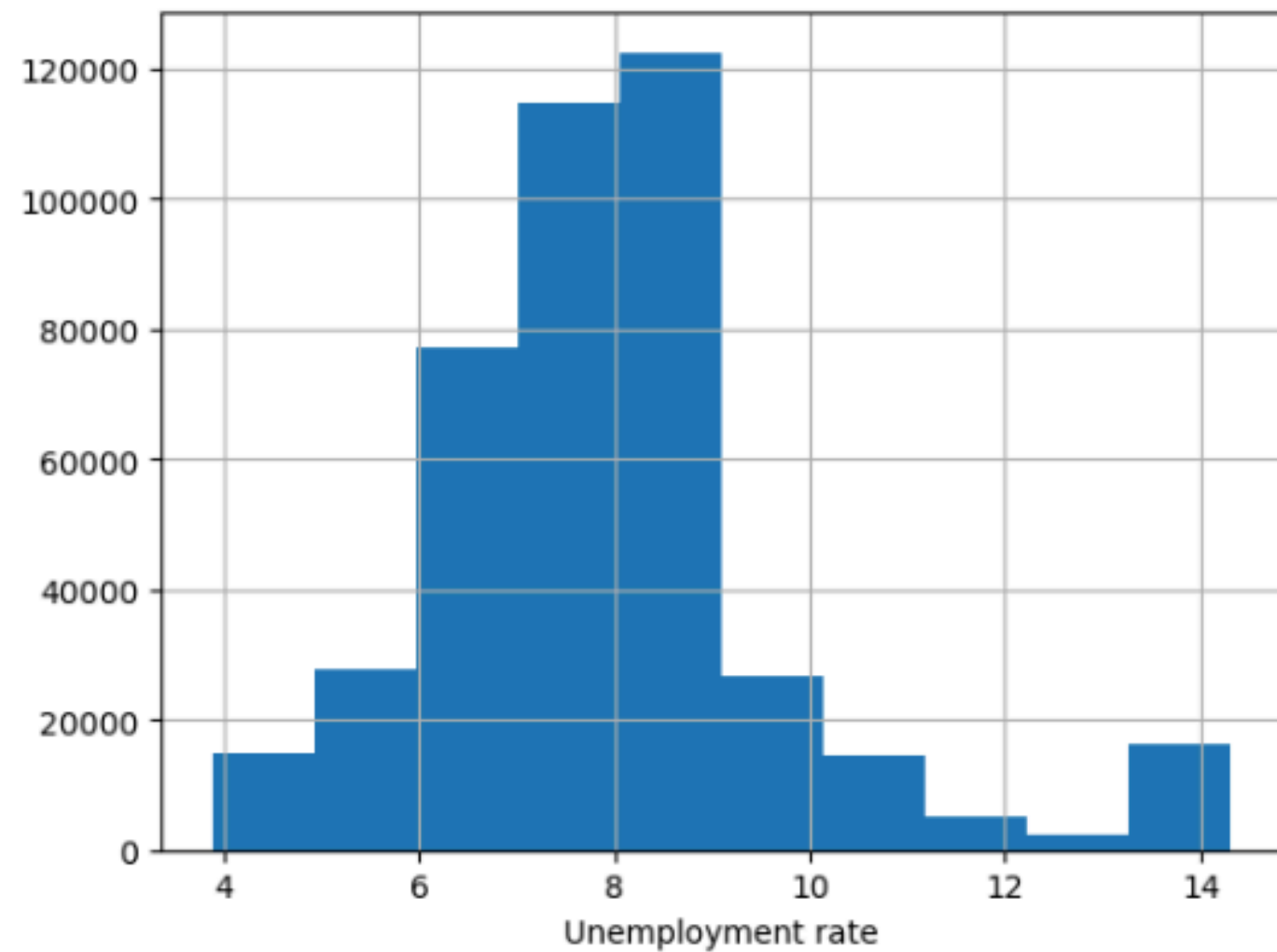


Vì 45 cửa hàng nằm ở 45 vùng khác nhau nên dựa vào đó ta có thể cơ bản biết được chỉ số CPI ở 45 vùng.

Khi thực hiện vẽ biểu đồ giữa chỉ số CPI và tổng doanh thu nhóm không thấy có được những mối quan hệ rõ ràng giữa 2 thuộc tính này

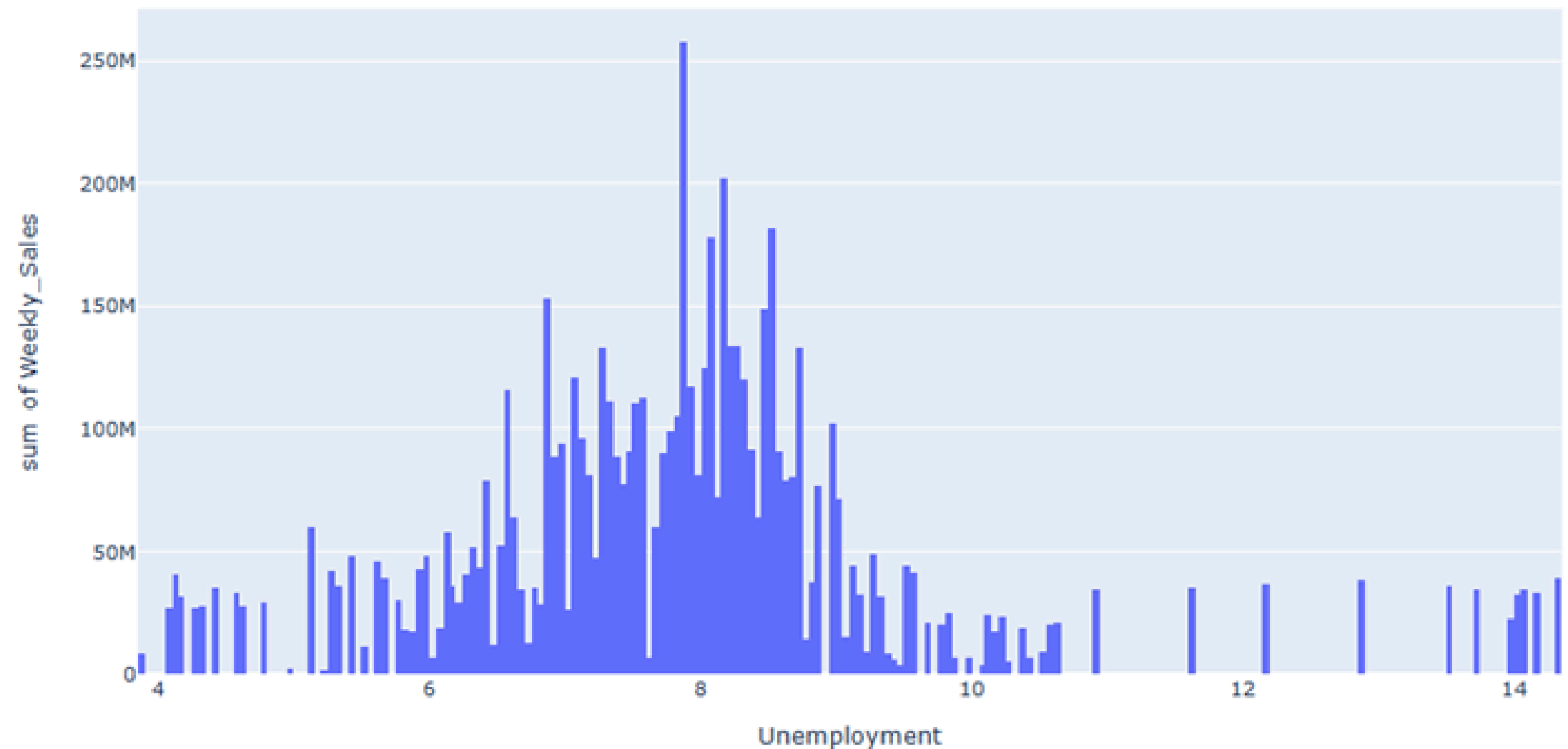


2.9 Tìm hiểu về tỷ lệ thất nghiệp.



Có thể thấy khi tỷ lệ thất nghiệp tăng cao như trong khoảng 10% đến 14% doanh thu đã bị giảm thể hiện tỷ lệ thất nghiệp cũng có ít nhiều ảnh hưởng đến doanh thu khi tỷ lệ thất nghiệp tăng, người tiêu dùng thường có xu hướng giảm chi tiêu do lo lắng về tình hình tài chính cá nhân và không chắc chắn về tương lai. Điều này có thể dẫn đến giảm doanh số bán hàng

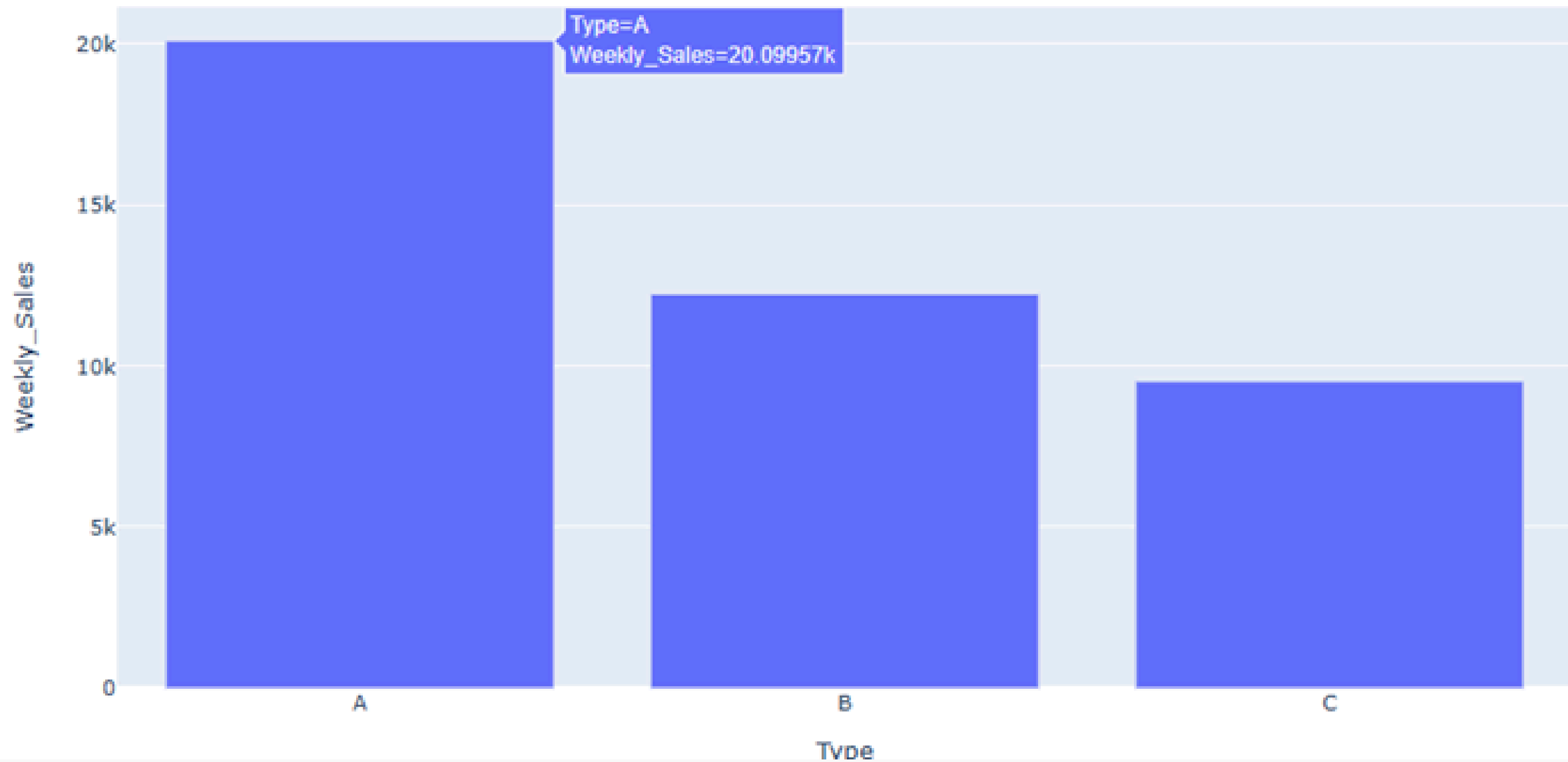
Qua biểu đồ trên ta thấy được rằng tỷ lệ thất nghiệp dao động chủ yếu ở mức 6 đến 10%



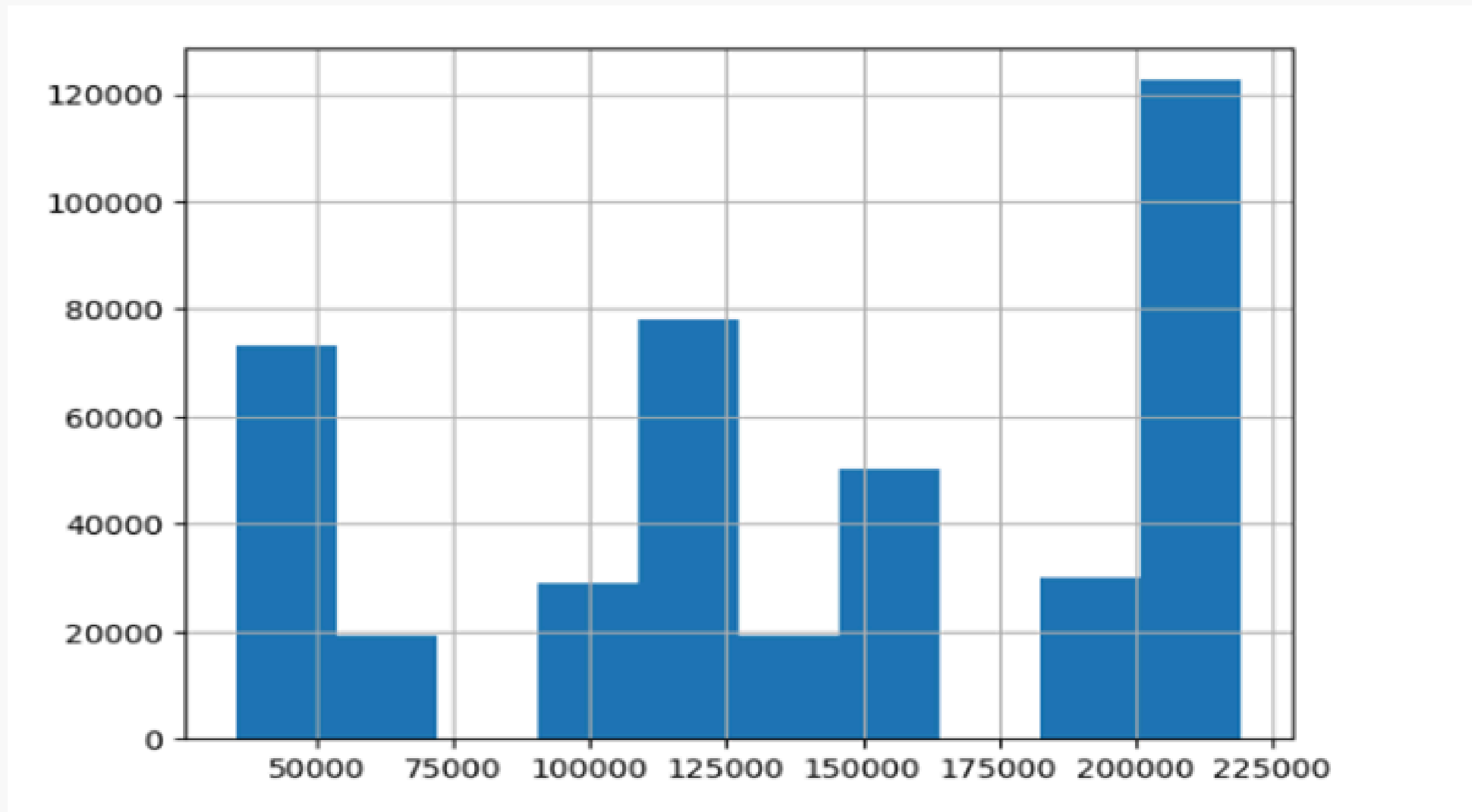
2.10 Tìm hiểu về thuộc tính Type

Có 3 loại cửa hàng trong bộ dữ liệu. Cửa hàng loại A chiếm đa số.

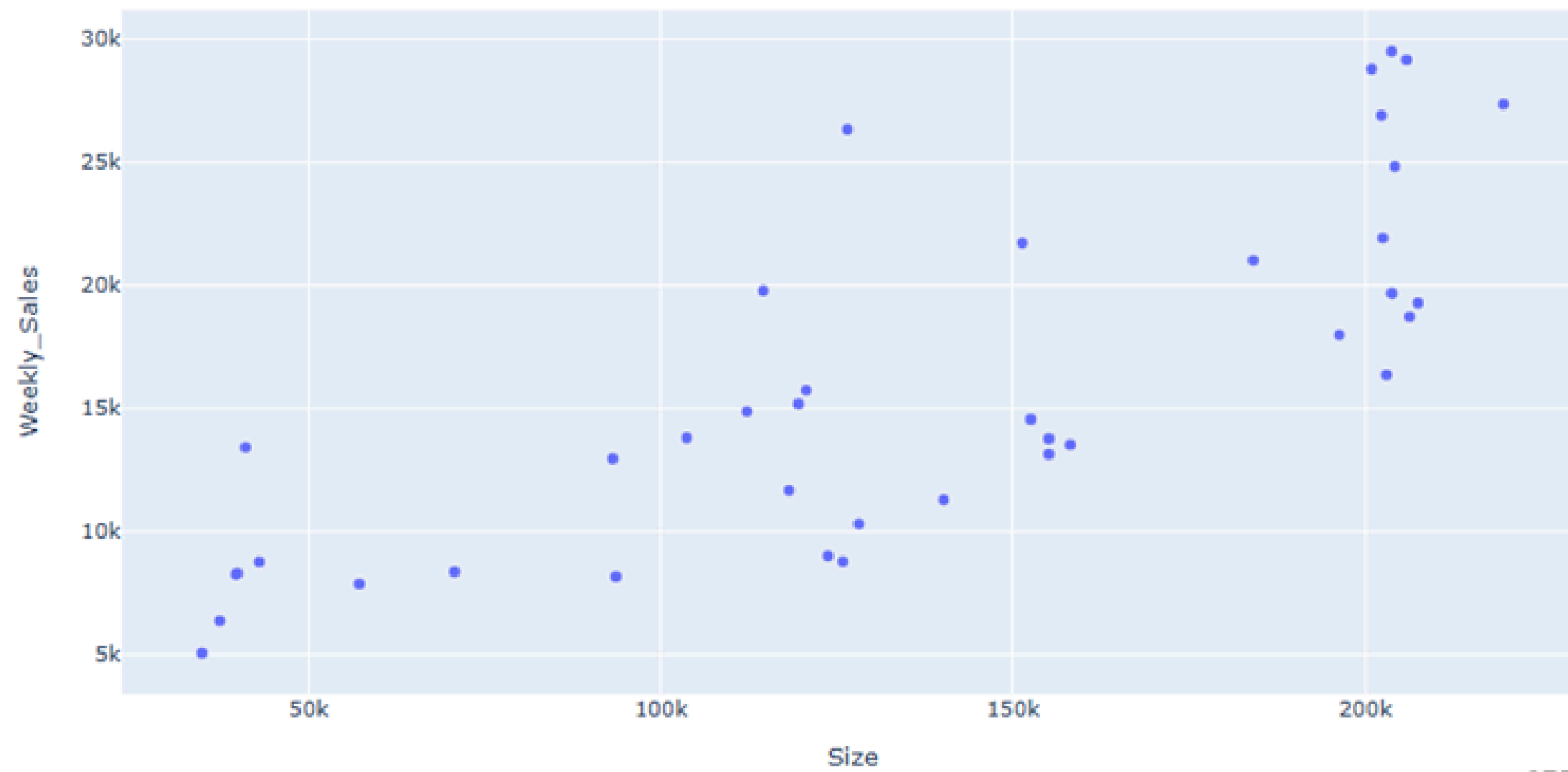
Trong 3 loại cửa hàng thì cửa hàng loại A cũng là cửa hàng có doanh thu trung bình cao nhất



2.11 Tìm hiểu về thuộc tính Size

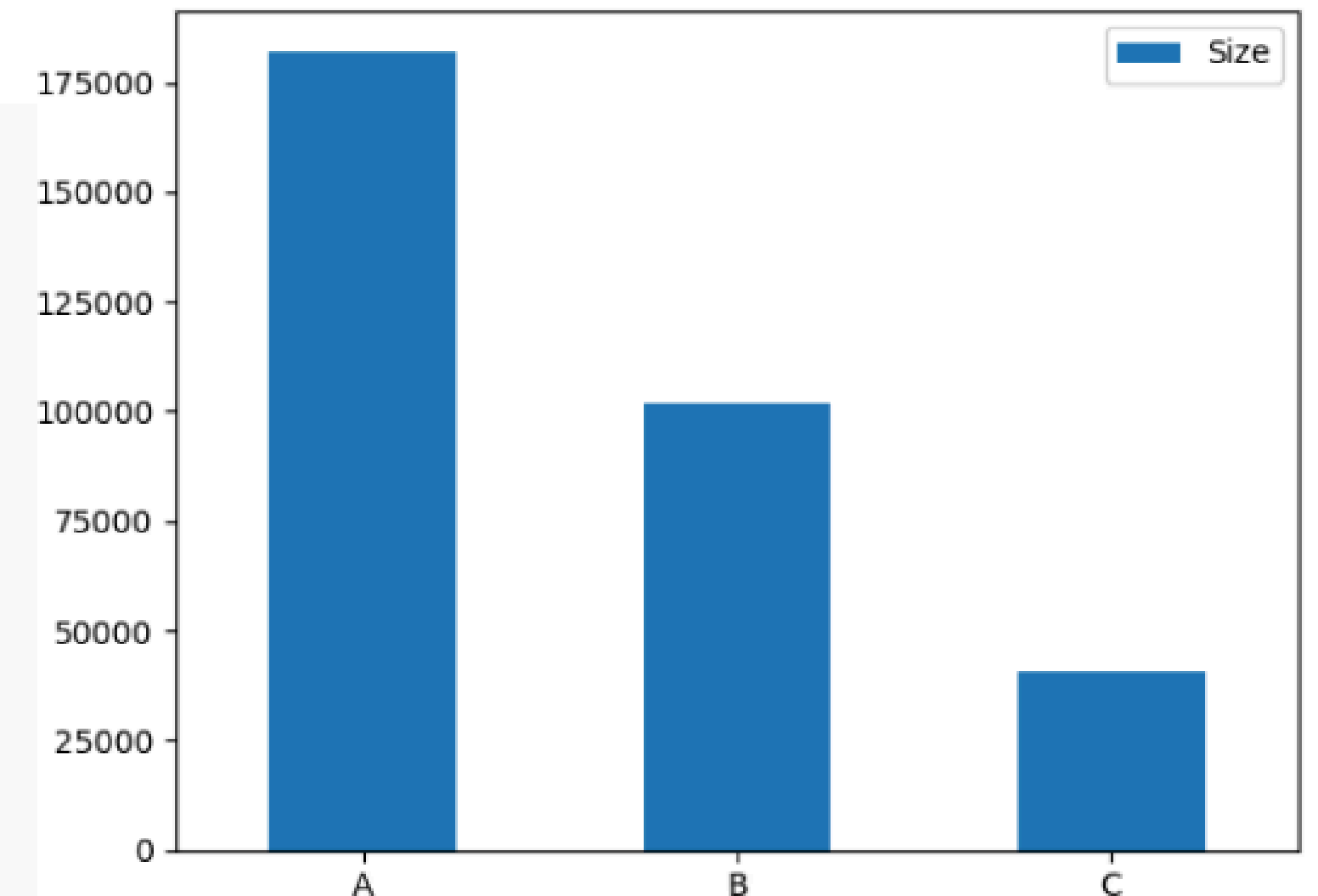


Có ba khoảng kích thước của cửa hàng từ 50000 -> 75000 từ 100000 -> 150000 và từ 200000



Có thể thấy được rằng kích cỡ của cửa hàng càng lớn thì doanh thu của cửa hàng càng cao.

Ta thấy được rằng store loại A cũng là store có kích thước lớn nhất cũng chứng minh cho nhận xét phía trên là store loại A có doanh thu cao nhất.



3. CHUẨN BỊ DỮ LIỆU (DATA PREPARATION)

Tập dữ liệu train sau khi đã được xử lý dữ liệu.

Thực hiện xử lý dữ liệu để chuẩn bị cho bước mô hình hóa

Do những dữ liệu markdown chứa quá nhiều dữ liệu Null nên nhóm đã thay thế bằng số 0

Trong tập dữ liệu test cũng xuất hiện dữ liệu Null trong thuộc tính CPI và Unemployment nên nhóm đã thực hiện lấp vào bằng giá trị trung bình của hai cột dữ liệu.

Với các dữ liệu boolean thể hiện cho các ngày lễ và dữ liệu thể hiện cho các loại cửa hàng nhóm sẽ thay đổi thành số bằng phương thức map của python.

Store	1	1	1	1	1
Dept	1	1	1	1	1
Date	2010-02-05	2010-02-12	2010-02-19	2010-02-26	2010-03-05
Weekly_Sales	24924.5	46039.49	41595.55	19403.54	21827.9
IsHoliday	0	1	0	0	0
Temperature	42.31	38.51	39.93	46.63	46.5
Fuel_Price	2.572	2.548	2.514	2.561	2.625
MarkDown1	0.0	0.0	0.0	0.0	0.0
MarkDown2	0.0	0.0	0.0	0.0	0.0
MarkDown3	0.0	0.0	0.0	0.0	0.0
MarkDown4	0.0	0.0	0.0	0.0	0.0
MarkDown5	0.0	0.0	0.0	0.0	0.0
CPI	211.096358	211.24217	211.289143	211.319643	211.350143
Unemployment	8.106	8.106	8.106	8.106	8.106
Type	1	1	1	1	1
Size	151315	151315	151315	151315	151315
Day	5	12	19	26	5
Week	5	6	7	8	9
Month	2	2	2	2	3
Year	2010	2010	2010	2010	2010

4 . MÔ HÌNH HÓA (MODELLING)

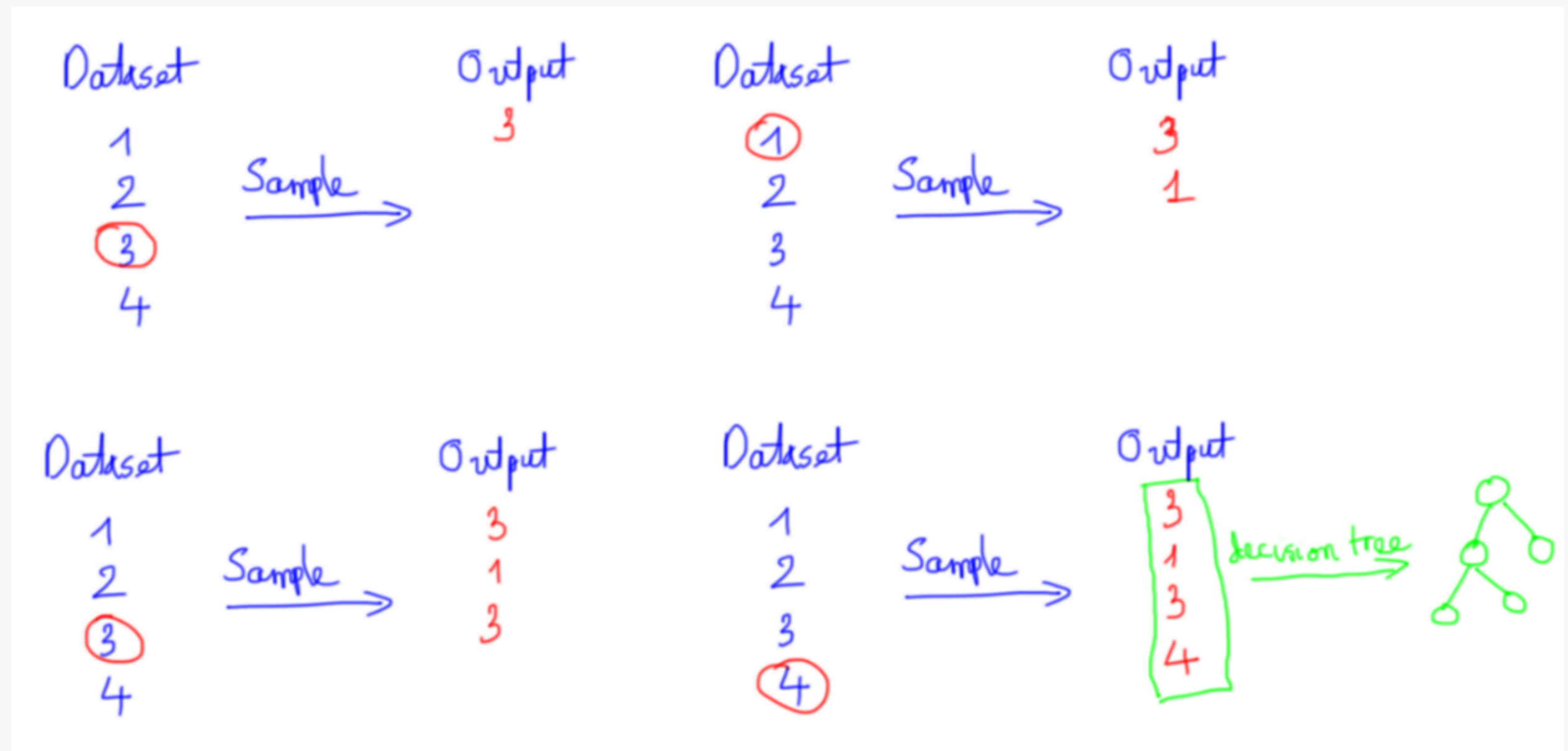
Thực hiện chia bộ dữ liệu train thành tập X và tập Y

Dùng train_test_splits để tạo ra bộ dữ liệu Train và Validation.

Thuật toán mà nhóm chọn để xây dựng mô hình là RandomForestRegressor

Giới thiệu về thuật toán Random Forest:

Thuật toán Random Forest là thuật toán sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.



5. ĐÁNH GIÁ (EVALUATION)

Kaggle cũng đã cung cấp công thức đánh giá hiệu suất nên nhóm đã tạo ra một phương thức đánh giá mô hình theo công thức WMAE (Weighted Mean Absolute Error) là một phương pháp đo lường sự chênh lệch giữa giá trị dự đoán và giá trị thực tế trong một mô hình học máy, với việc áp dụng trọng số khác nhau cho các mẫu dữ liệu khác nhau. WMAE thường được sử dụng trong các bài toán dự báo.

This competition is evaluated on the weighted mean absolute error (WMAE):

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^n w_i |y_i - \hat{y}_i|$$

where

- n is the number of rows
- \hat{y}_i is the predicted sales
- y_i is the actual sales
- w_i are weights. $w = 5$ if the week is a holiday week, 1 otherwise

Số điểm của mô hình khi được đánh giá trên bộ dữ liệu Validation

```
y_preds_1 = rfr.predict(X_valid)
```

```
WMAE(X_valid,y_valid,y_preds_1)
```

1526.12

Kết quả thứ 2 là mô hình được huấn luyện trên bộ dữ liệu đã được xóa các cột có độ tương quan thấp với doanh thu.

Kết quả đầu tiên là mô hình được huấn luyện trên bộ dữ liệu với đầy đủ các cột.

```
ypreds = rfr.predict(X_valid)
```

```
WMAE(X_valid,y_valid,ypreds)
```



1426.23



6. DEPLOY

Sao khi thực hiện đánh giá model nhóm nhận thấy đây là hai mô hình ưng ý nên quyết định thực hiện cho mô hình thực hiện dự đoán trên tập dữ liệu X_test đã được Kaggle cung cấp sao đó nộp file dự đoán của hai mô hình để trang web Kaggle thực hiện đánh giá.

Kết quả của hai mô hình được đánh giá bởi Kaggle:

Submission and Description		Private Score ⓘ	Public Score ⓘ	Selected
	final2.csv Complete (after deadline) · 5h ago · kaggle competitions submit -c walmart-recruiting-store-sales...	2858.74373	2779.78556	<input type="checkbox"/>
	final1.csv Complete (after deadline) · 5h ago · My final for this exam	3797.90884	3651.50021	<input type="checkbox"/>





Thank you

