Store Sales Forecasting

Done by:

Hamza Naser

Diaa Alqadi

Dana Ghazal

Problem Definition

Basic question to be resolved

How to increase store sales and attract more store customers during the next 3 months?

Initial situation

The store is experiencing sales that are not meeting desired targets or growth expectations. The store operates in a competitive retail environment. Over the next 3 months, achieve a 5% growth in the number of customers.

Decision makers

- Store Manager
- Regional Sales Director
- Marketing Manager

Decision/success criteria

- Increase weekly sales by 10% through the establishment of new branches across various regions.
- Improve customer engagement and enhance their experience.
- Improve marketing and sales strategies.

Constraints

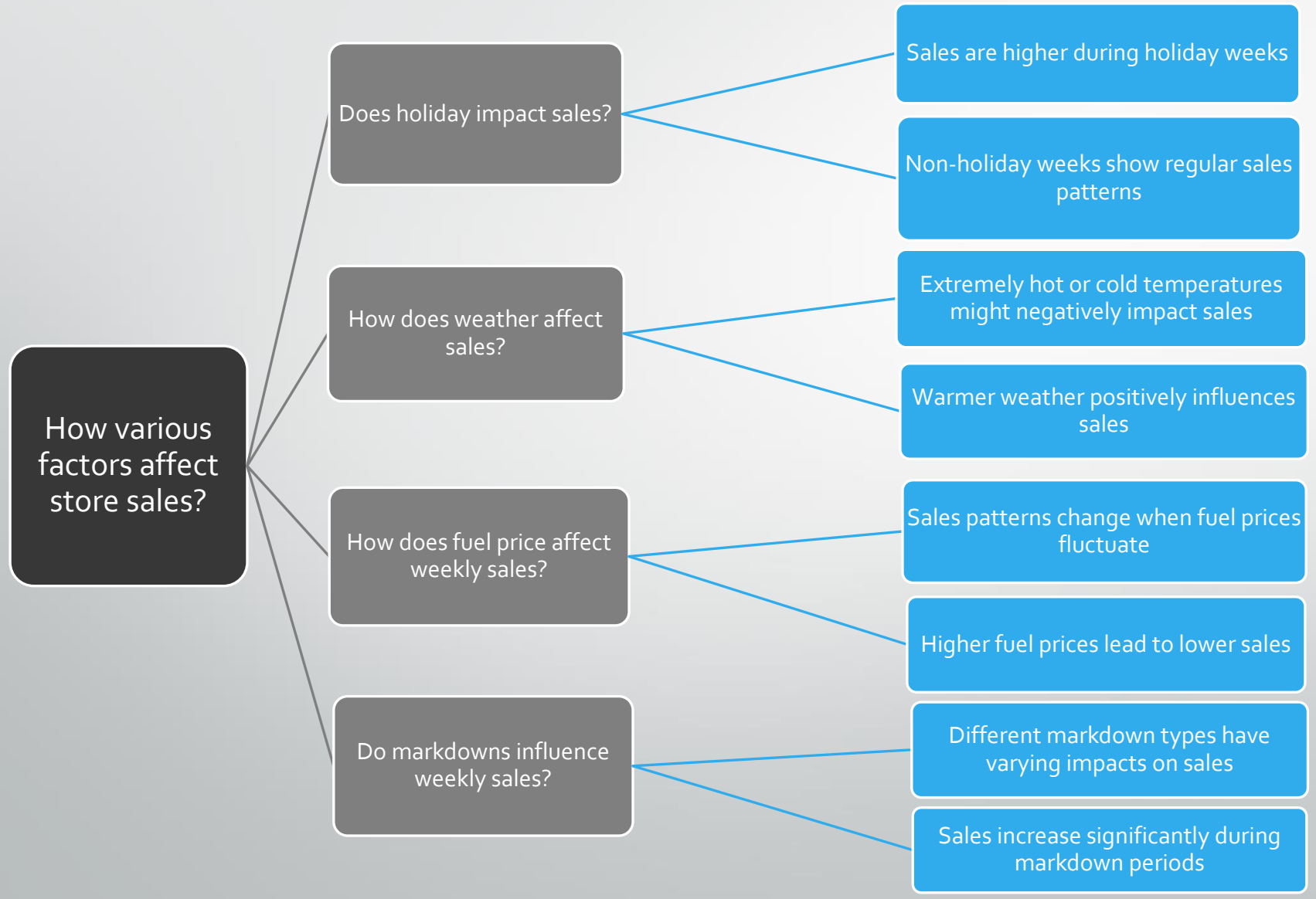
- Budget limitations for opening new branches and for implementing sales improvement strategies.
- Compliance with legal and ethical standards in marketing and sales activities.

Project scope

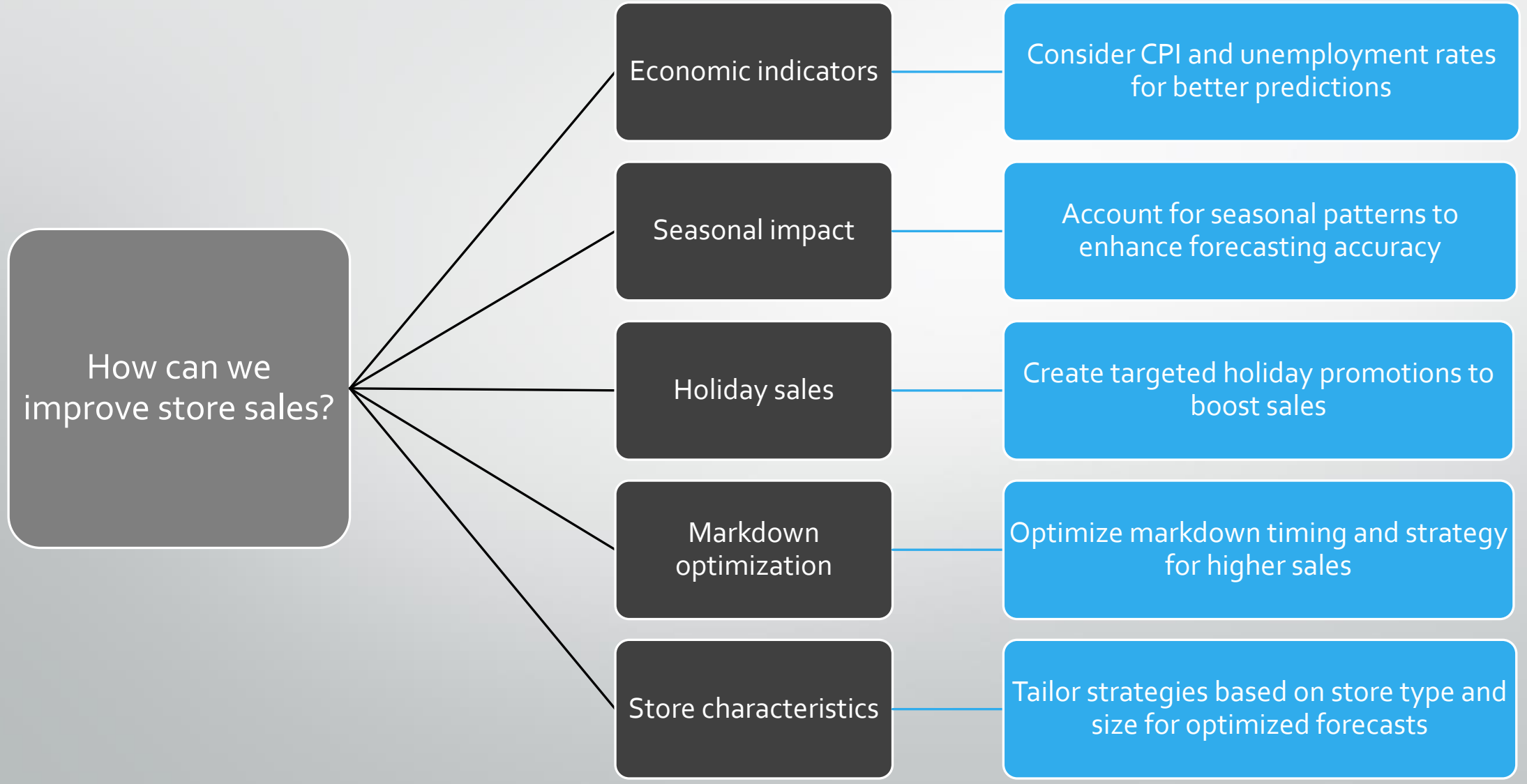
Focus on the 45 stores and on the regions that achieve the largest number of sales.

Problem Structuring

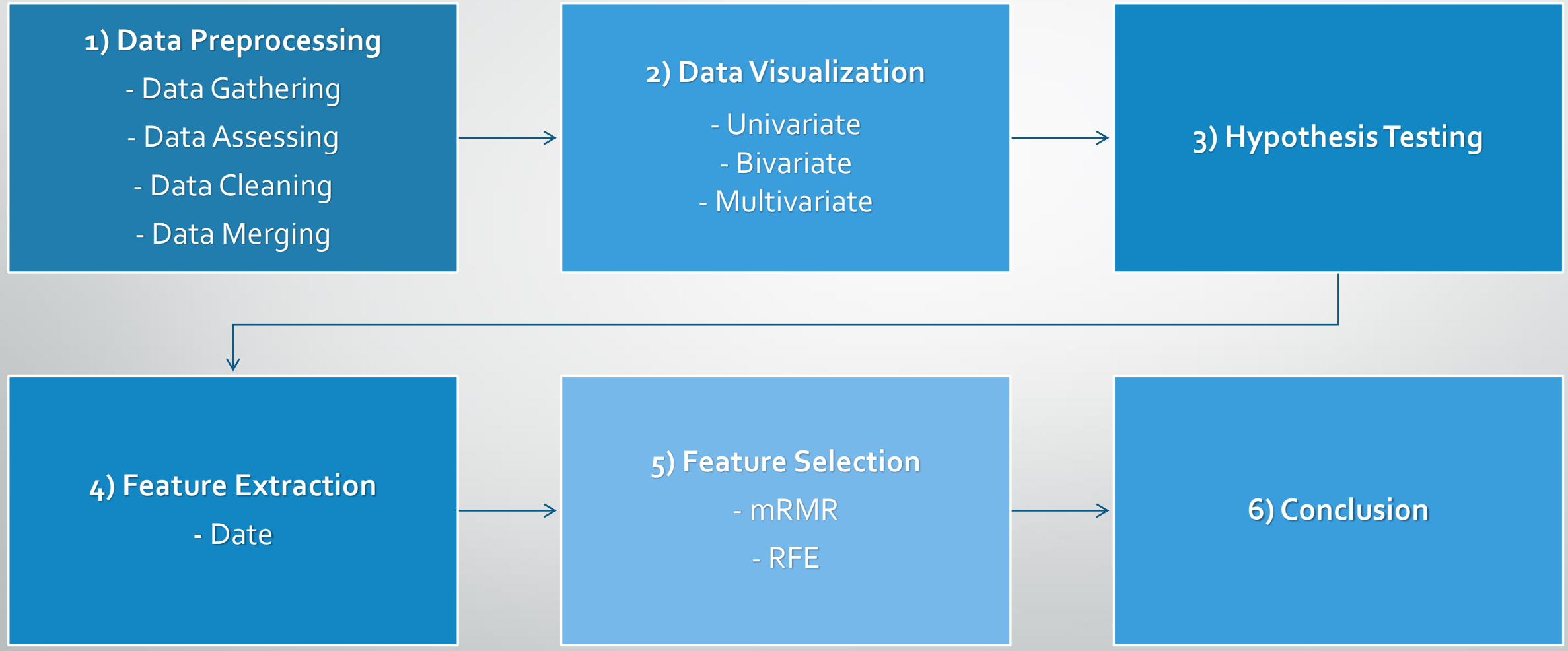
Hypothesis Tree



Deductive Tree



Tasks Identification


1) Data Preprocessing
- Data Gathering
- Data Assessing
- Data Cleaning
- Data Merging
2) Data Visualization
- Univariate
- Bivariate
- Multivariate
3) Hypothesis Testing
4) Feature Extraction
- Date
5) Feature Selection
- mRMR
- RFE
6) Conclusion

1) Data Preprocessing

- **Data Gathering**

Our dataset includes 3 tables where provided as .CSV file format that contains information about sales for 45 Walmart stores located in different regions.

Stores Dataset: General informations about each store, 45 rows x 3 columns

Column	Description	Data Type
Store	Stores numbers from 1 to 45	Categorical
Type	Store type has been provided, there are 3 types A, B, and C	Categorical
Size	Stores size	Numerical

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

Features Dataset: Records store-specific data for different business weeks, categorized into Regional Information (like Temperature and Unemployment) and Promotional Information (such as Advertising options). 8,190 rows x 12 columns

Column	Description	Data Type
Store	The store number	Categorical
Date	Day of the week	Categorical
Temperature	Average temperature in the region in Fahrenheit	Numerical
Fuel_Price	Cost of fuel in the region in Dollars	Numerical
MarkDown1	Anonymized data related to promotional markdowns that Walmart is running	Numerical
MarkDown2	Anonymized data related to promotional markdowns that Walmart is running	Numerical
MarkDown3	Anonymized data related to promotional markdowns that Walmart is running	Numerical
MarkDown4	Anonymized data related to promotional markdowns that Walmart is running	Numerical
MarkDown5	Anonymized data related to promotional markdowns that Walmart is running	Numerical
CPI	The consumer price index	Numerical
Unemployment	The unemployment rate	Numerical
IsHoliday	Whether the week is a special holiday week	Categorical

	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

Train Dataset: Weekly Sales for each store/dept, 421,570 rows x 5 columns

Column	Description	Data Type
Store	The store number	Categorical
Dept	The department number	Categorical
Date	Day of the week	Categorical
Weekly_Sales	Sales for the given department in the given store in Dollars	Numerical
IsHoliday	Whether the week is a special holiday week	Categorical

	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

Data Assessing

- **Checking for NULL Values:**

We noticed that features table contains null values in below columns:

- The five MarkDown columns (1-5)
- CPI and Unemployment columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8190 entries, 0 to 8189
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         8190 non-null   int64
 1   Date          8190 non-null   object
 2   Temperature   8190 non-null   float64
 3   Fuel_Price    8190 non-null   float64
 4   MarkDown1     4032 non-null   float64
 5   MarkDown2     2921 non-null   float64
 6   MarkDown3     3613 non-null   float64
 7   MarkDown4     3464 non-null   float64
 8   MarkDown5     4050 non-null   float64
 9   CPI           7605 non-null   float64
 10  Unemployment  7605 non-null   float64
 11  IsHoliday     8190 non-null   bool
dtypes: bool(1), float64(9), int64(1), object(1)
memory usage: 712.0+ KB
```

- **Checking for -ve Values:**

We noticed that we got –ve values in features table in Markdown columns (1-3) and 5.

Out[89]:

	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5
min	-2781.45	-265.76	-179.26	0.22	-185.17

We also got 1285 –ve values in Weekly_Sales column in the train table.

Out[95]:

	Weekly_Sales
min	-4988.94

- **Checking for Data Type:**

Data type is wrong for below columns:

Column	Type	Correct Type
Store	int64	category
Dept	int64	category
Date	object	datetime64
Type	object	category

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421570 entries, 0 to 421569
Data columns (total 5 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   Store         421570 non-null  int64
 1   Dept          421570 non-null  int64
 2   Date          421570 non-null  object
 3   Weekly_Sales  421570 non-null  float64
 4   IsHoliday     421570 non-null  bool
dtypes: bool(1), float64(1), int64(2), object(1)
memory usage: 13.3+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45 entries, 0 to 44
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Store   45 non-null     int64
 1   Type    45 non-null     object
 2   Size    45 non-null     int64
dtypes: int64(2), object(1)
memory usage: 1.2+ KB
```

- **Checking for Outliers:**

We got outliers values in six columns as shown below:

- Weekly_Sales
- The five MarkDown columns (1-5)

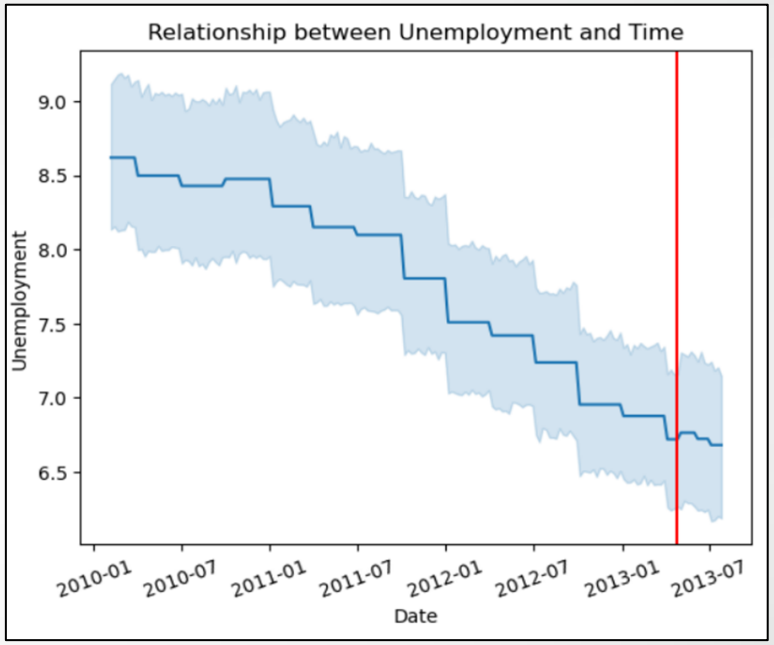
```
The number of outliers of the column Weekly_Sales is 35521
The number of outliers of the column MarkDown1 is 237
The number of outliers of the column MarkDown2 is 436
The number of outliers of the column MarkDown3 is 480
The number of outliers of the column MarkDown4 is 337
The number of outliers of the column MarkDown5 is 212
```

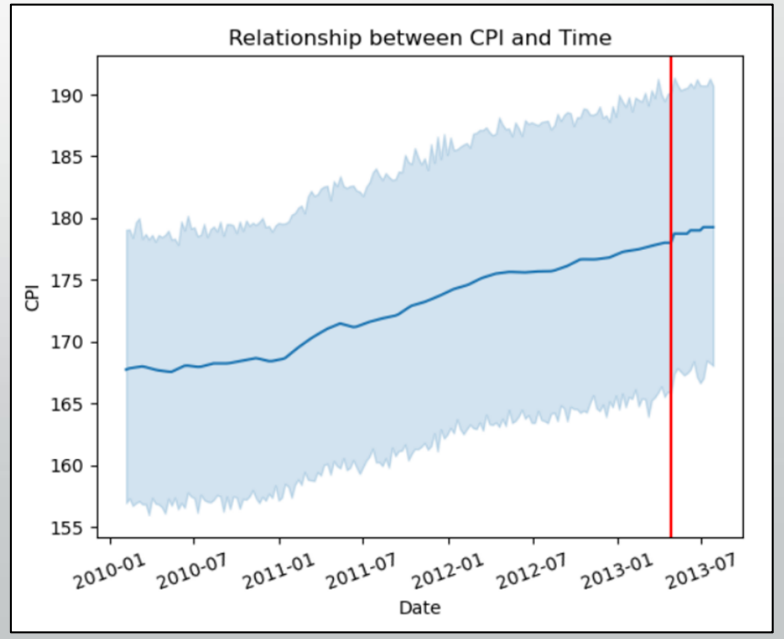
- # Data Cleaning

  - **Handling Missing Values:**

Nulls in markdown columns imply no promotions for those weeks, and we have addressed them by replacing them with zeros.

CPI and Unemployment columns contain nulls influenced by time (Date) and location (Store number). To tackle this, we built a linear regression model to impute and fill nulls in CPI and Unemployment based on time and store.





- **Correct the Data Type:**

We have changed the data type in the tables for the column Date to date time, Store to category, Dept to category, and Type to category.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421570 entries, 0 to 421569
Data columns (total 5 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   Store         421570 non-null  category
 1   Dept          421570 non-null  category
 2   Date          421570 non-null  datetime64[ns]
 3   Weekly_Sales  421570 non-null  float64
 4   IsHoliday     421570 non-null  bool
dtypes: bool(1), category(2), datetime64[ns](1), float64(1)
memory usage: 7.6 MB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45 entries, 0 to 44
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Store   45 non-null     category
 1   Type    45 non-null     category
 2   Size    45 non-null     int64
dtypes: category(2), int64(1)
memory usage: 2.1 KB
```

- **Drop the Negative Values:**

We have dropped rows in the feature table for Markdown columns (1-3) and 5 that have negative values.

We also dropped rows in the train table for Weekly_Sales column which contain negative values.

- **Data Merging**

We have merged the three distinct tables, namely the Features table, Stores table, and Train table, into a consolidated and comprehensive master dataset.

This unified dataset will contain a harmonized representation of store-specific information, including regional and promotional details, sales data, and store attributes.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 418727 entries, 0 to 418726
Data columns (total 16 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   Store         418727 non-null  category
 1   Dept          418727 non-null  category
 2   Date          418727 non-null  datetime64[ns]
 3   Weekly_Sales  418727 non-null  float64
 4   IsHoliday     418727 non-null  bool
 5   Type          418727 non-null  category
 6   Size          418727 non-null  int64
 7   Temperature   418727 non-null  float64
 8   Fuel_Price    418727 non-null  float64
 9   MarkDown1     418727 non-null  float64
 10  MarkDown2     418727 non-null  float64
 11  MarkDown3     418727 non-null  float64
 12  MarkDown4     418727 non-null  float64
 13  MarkDown5     418727 non-null  float64
 14  CPI           418727 non-null  float64
 15  Unemployment  418727 non-null  float64
dtypes: bool(1), category(3), datetime64[ns](1), float64(10), int64(1)
memory usage: 43.1 MB
```

- Since we are trying to predict weekly sales its not a wise move to drop outliers since this will affect the model's performance and the ability to predict the sales for out-of-sample data. The same thing applies for Mark Down Promotions.

- We converted the Temperature column values from Fahrenheit to Celsius to make it more interpretable.

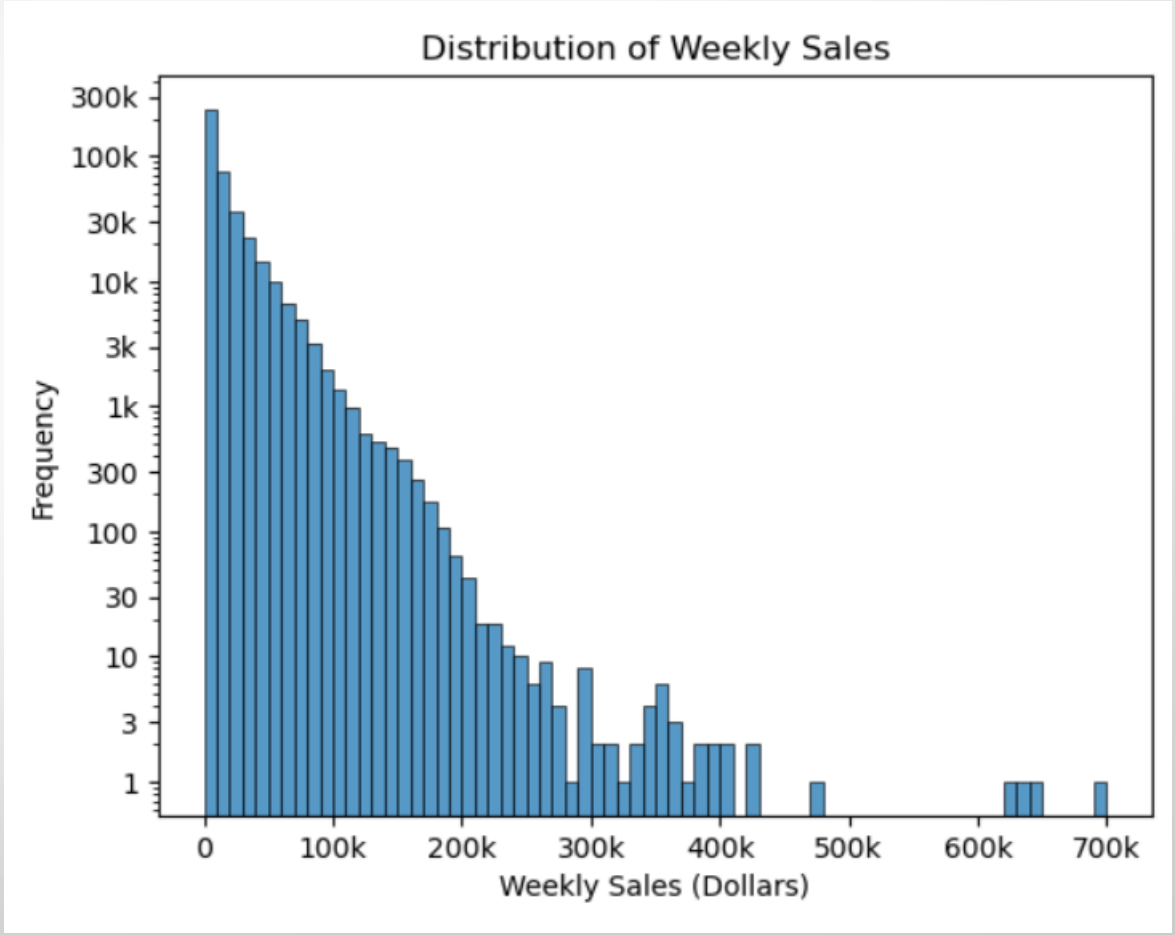
2) Data Visualization

Univariate Visualizations

What is the distribution of weekly sales?

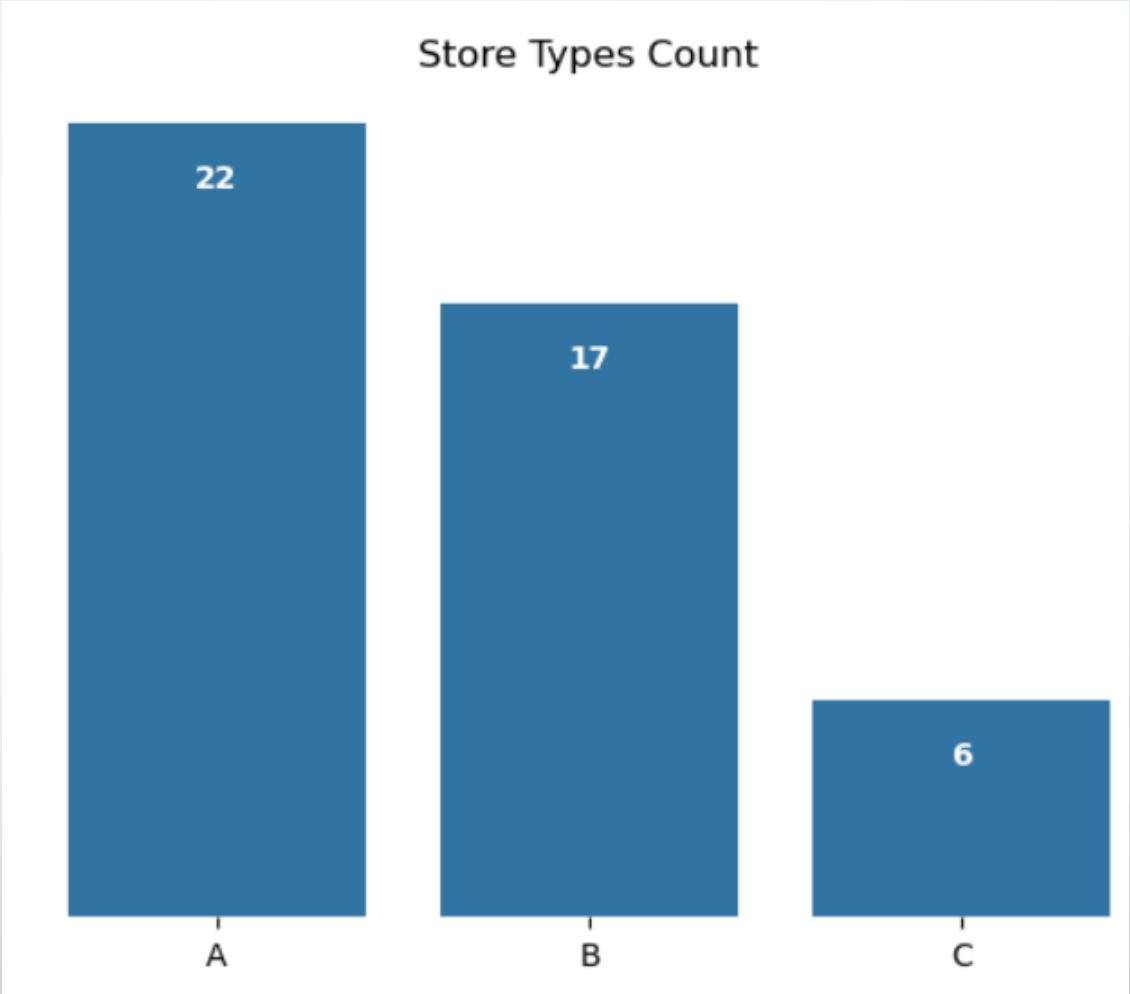
The distribution of weekly sales is highly skewed to the right.

The most common weekly sales are between $0 and $20k.



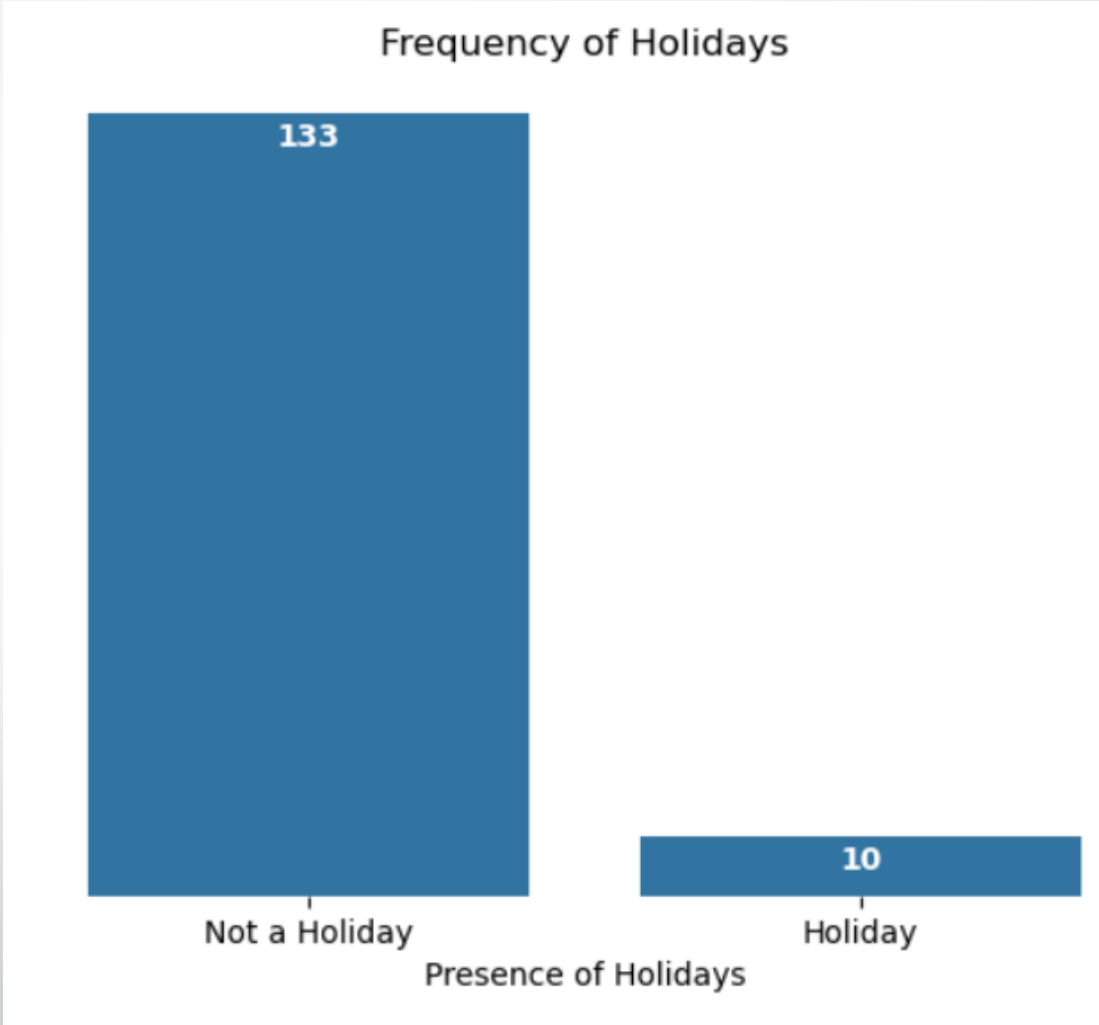
How many distinct store types are present in our dataset?

The dataset appears to have three types of stores, type A being the most popular and type C being the least popular.



How many holidays are included in the set of unique dates within the dataset?

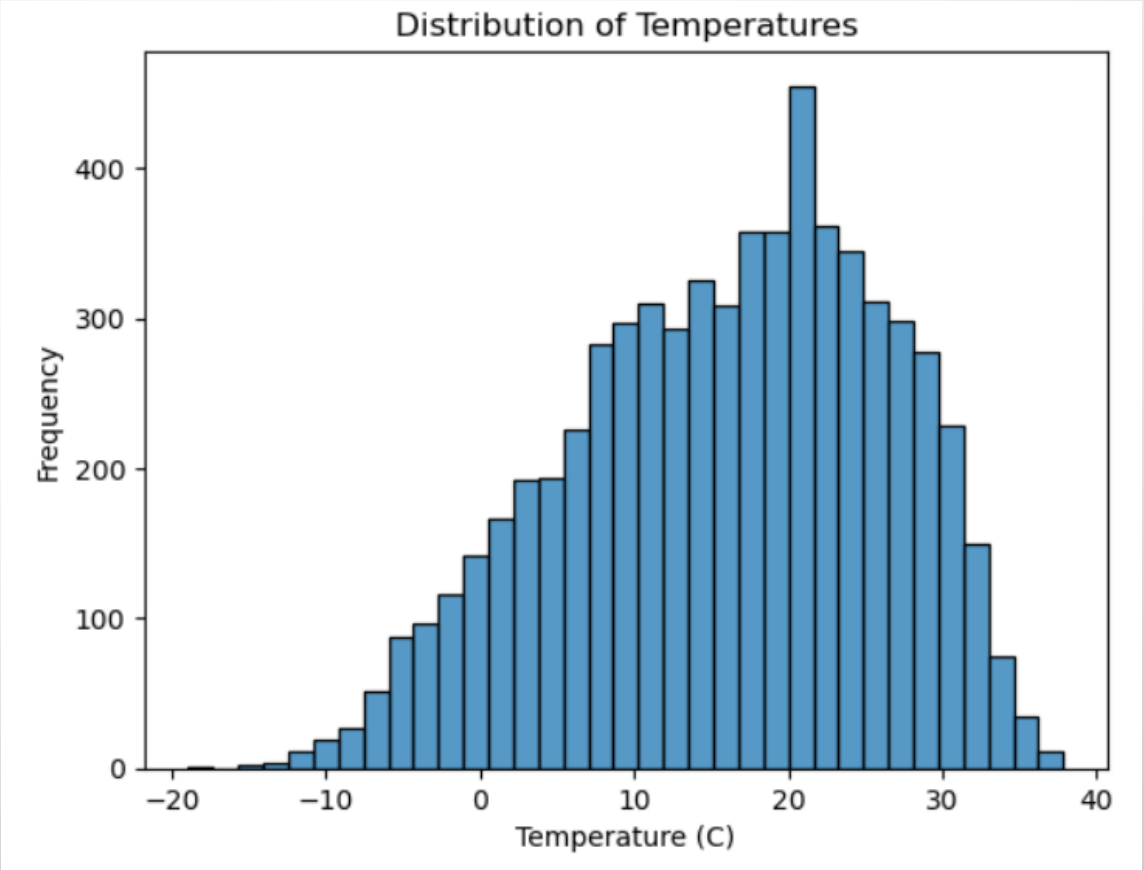
During the period from Feb 5, 2010, to Oct 26, 2012, a total of 10 weeks featured holidays, while 133 weeks passed without holidays.



How are the temperatures distributed over the weeks?

The temperature distribution is approximately normal.

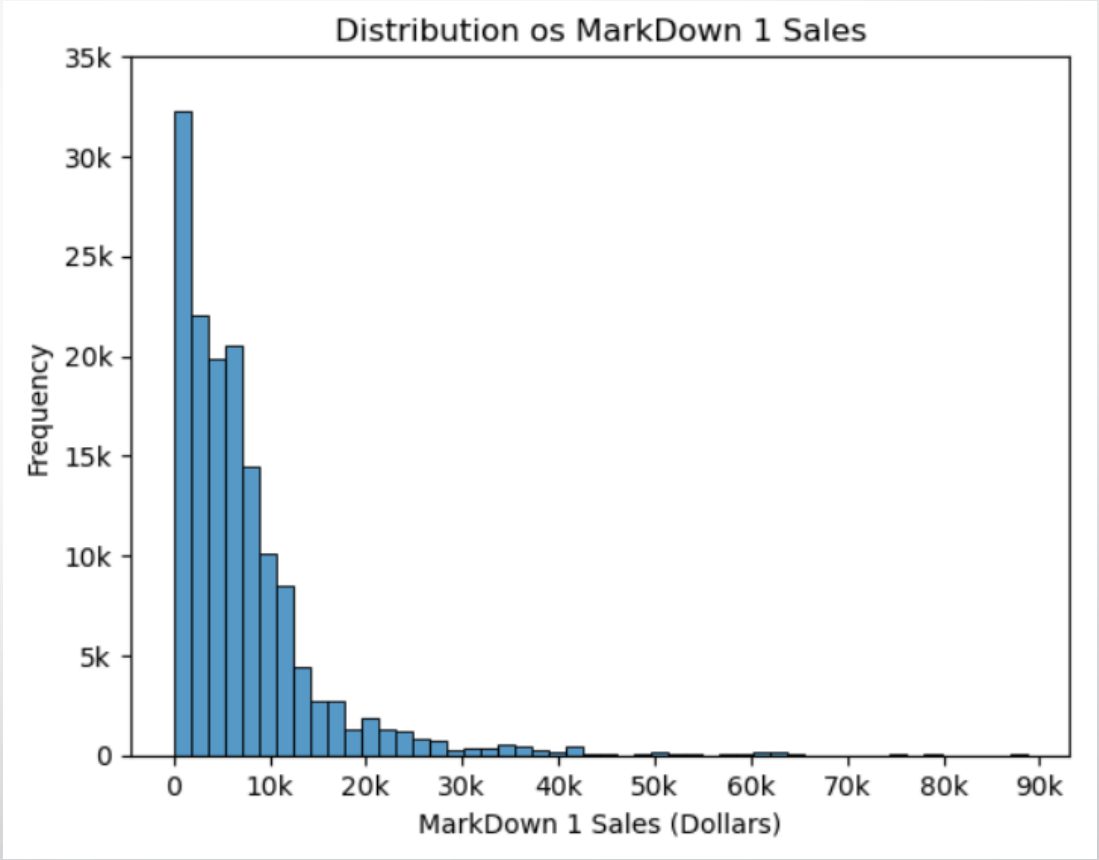
The most common temperatures are between 15°C and 25°C.



What is the arrangement of sales across Markdown 1?

The distribution of markdown 1 sales is highly skewed to the right.

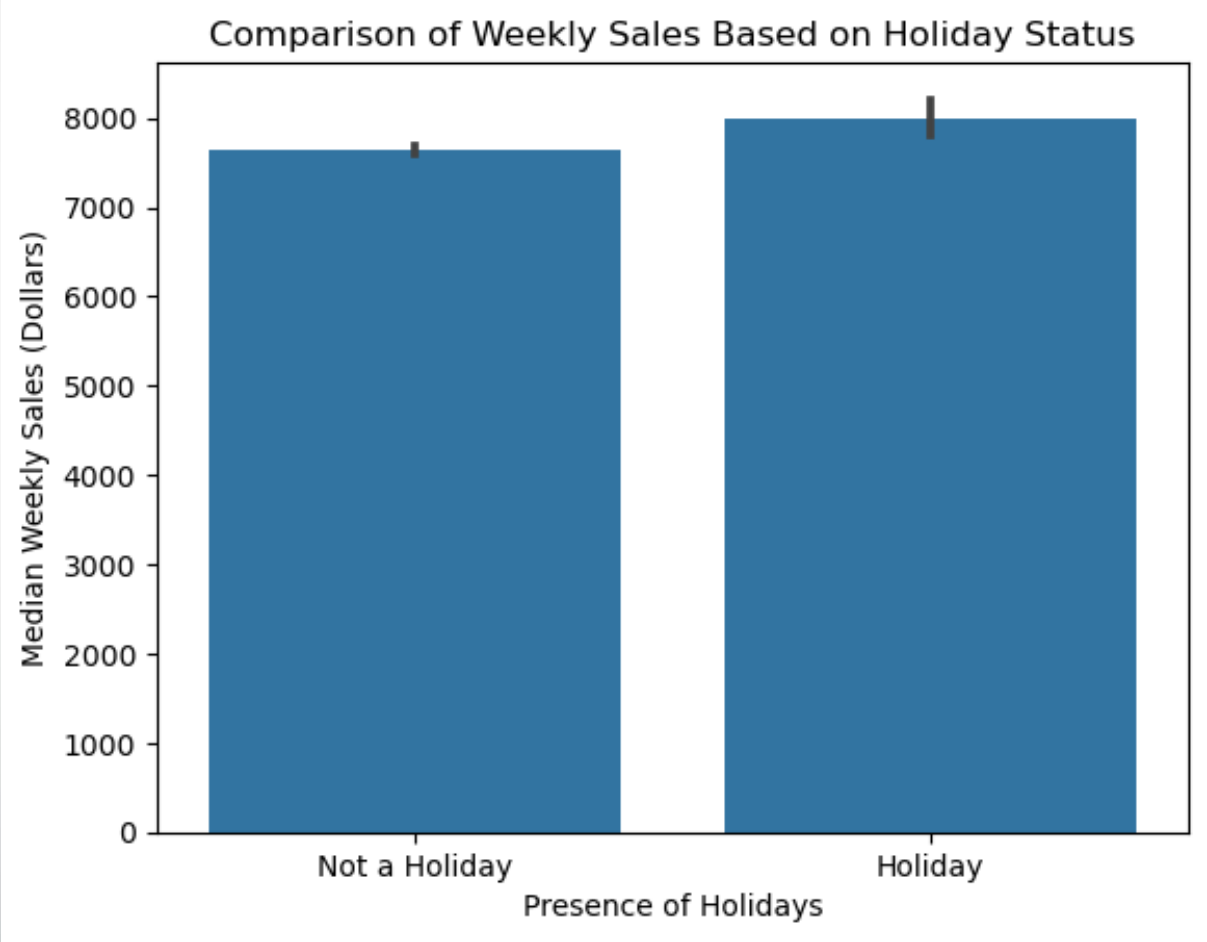
The most common amount of markdown 1 sales is between $0 and $10k.



Bivariate Visualizations

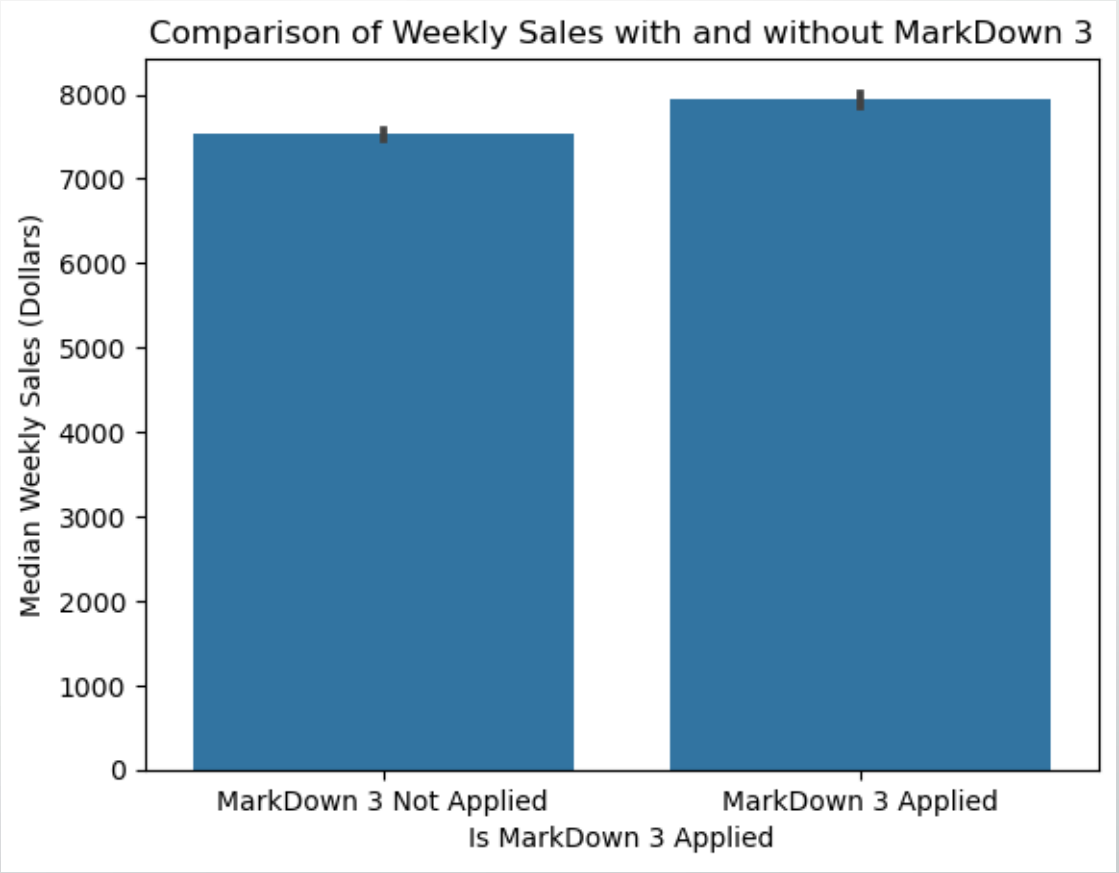
Is there a noticeable trend in weekly sales in relation to holiday and non-holiday weeks?

Sales show a minor increase during holiday weeks compared to regular weeks, forming the basis for a hypothesis we aim to test for statistical significance.



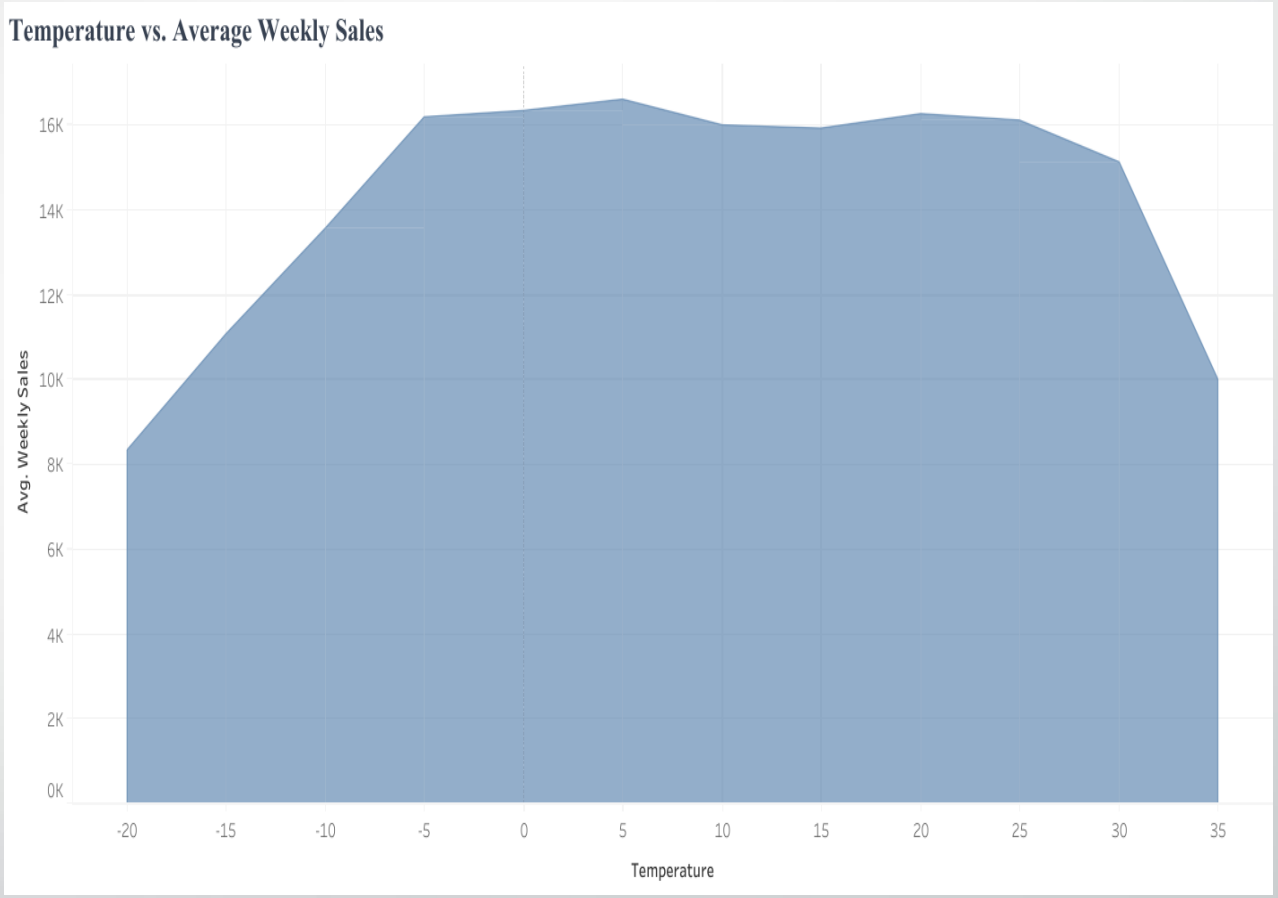
Does the presence of Markdown 3 correlate with observable differences in weekly sales behavior?

Using Markdown 3 leads to higher weekly sales, implying more purchases, but more analysis is needed to confirm this and consider other factors.



Does the weather conditions affect weekly sales?

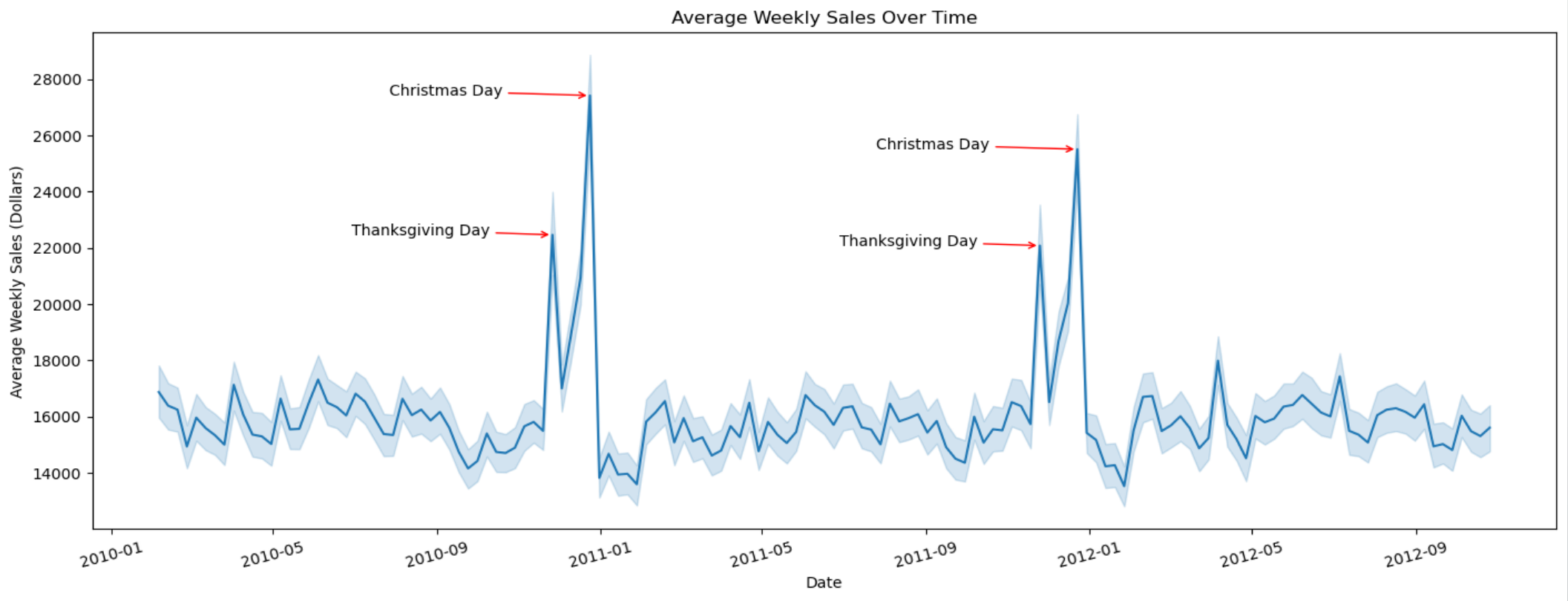
Weekly Sales are affected only on extreme high, or extreme low temperatures.



How does the median size of Walmart stores vary across different store types?

Among Walmart store types A, B, and C, type A is the biggest on average and has the most size variation, suggesting a mix of large and small stores with diverse strategies.

How do holidays like Christmas Day and Thanksgiving Day impact the pattern of average weekly sales over the observed period?

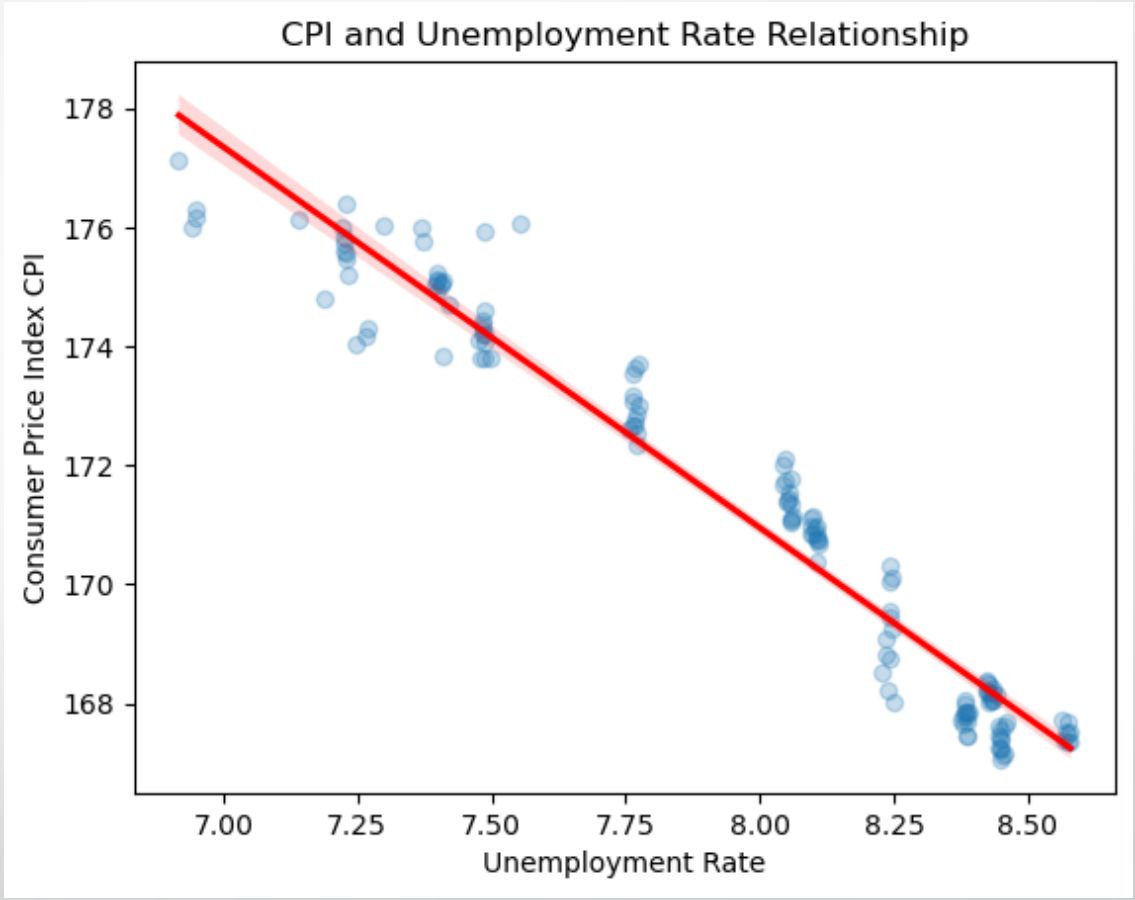


Holidays like Christmas and Thanksgiving have the highest average weekly sales, showing that people spend more during these times.

Is there a discernible relationship or correlation between the Consumer Price Index (CPI) and the Unemployment Rate?

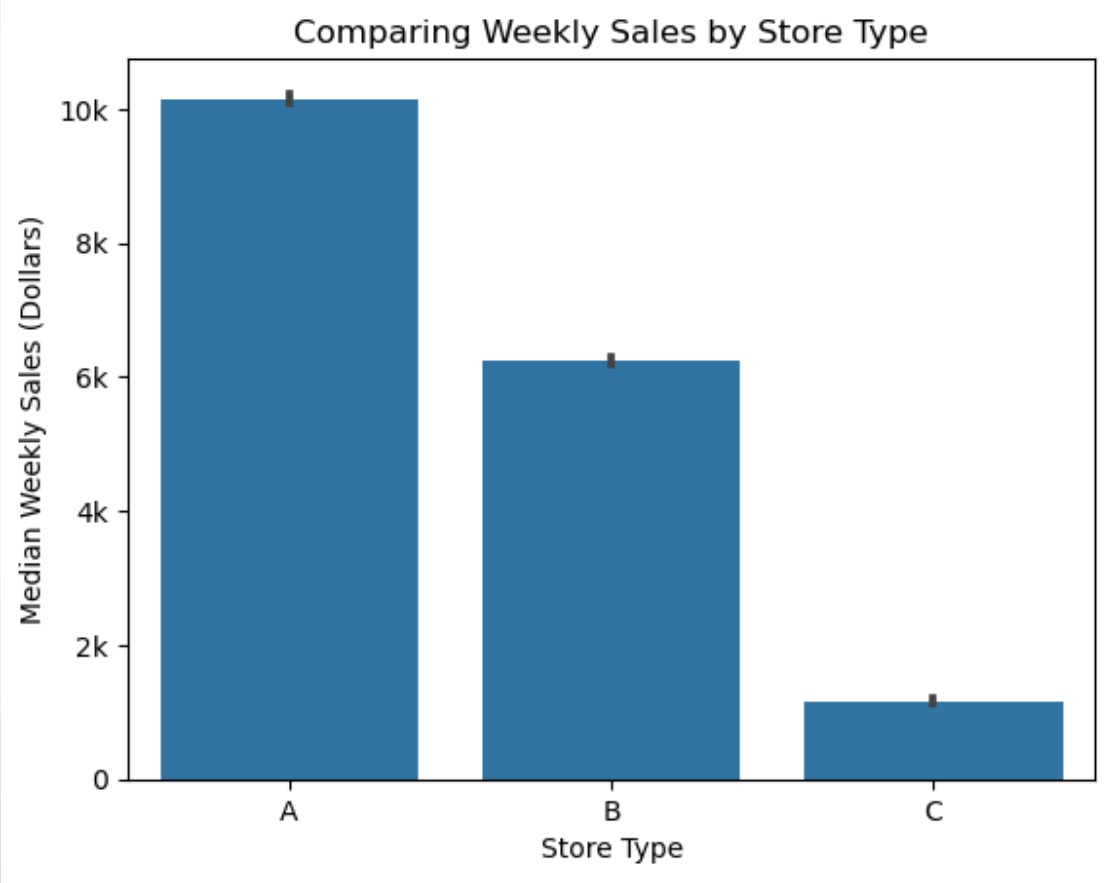
1- In economics, when unemployment drops, inflation tends to rise. This is called the Phillips Curve.

2- Consumer Price Index (CPI) up, prices up; CPI down, prices down. High CPI, high inflation; low CPI, low inflation or deflation.



Which store type demonstrates the highest median weekly sales, and how do other types compare?

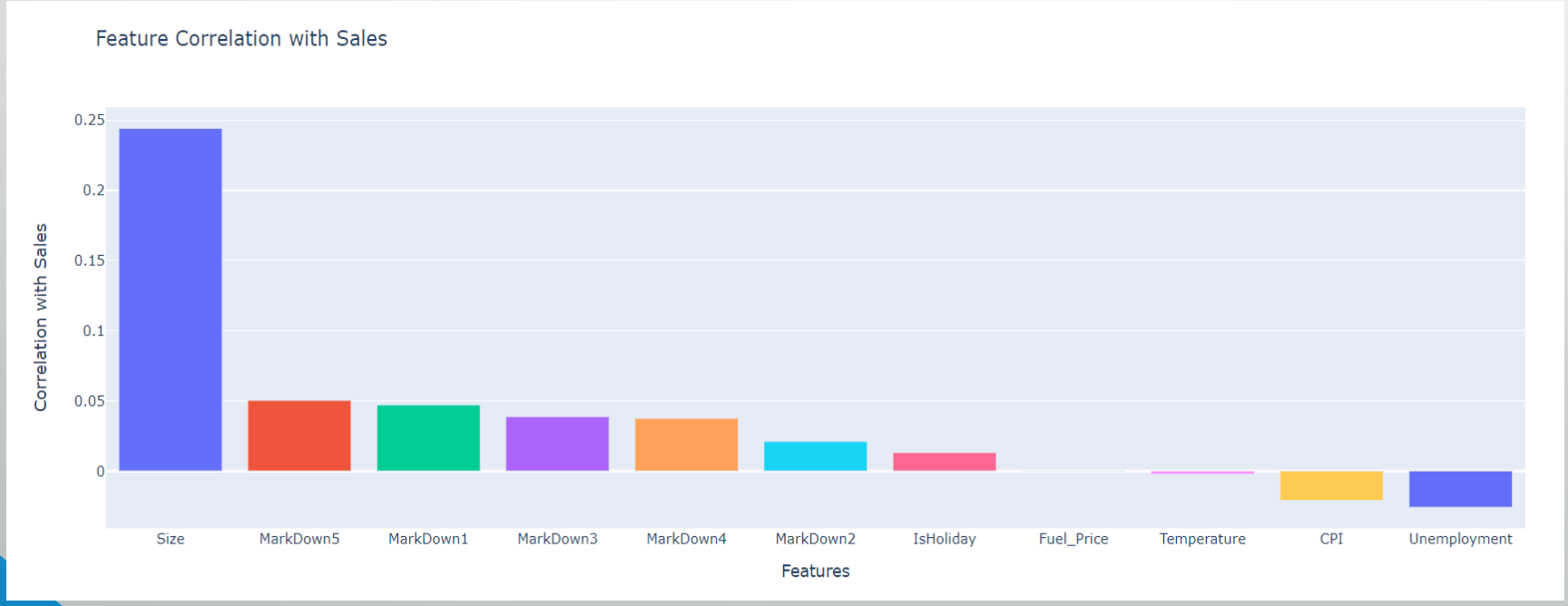
Store type A has the highest median weekly sales among various types. It consistently outperforms types B and C in terms of weekly sales, indicating its better performance.



How much other variables correlated with weekly sales?

As shown below, Size is highly correlated with Sales, as larger stores are expected to generate higher sales.

Also, one thing to notice is that fuel price almost no effect on sales based on correlation point of view.

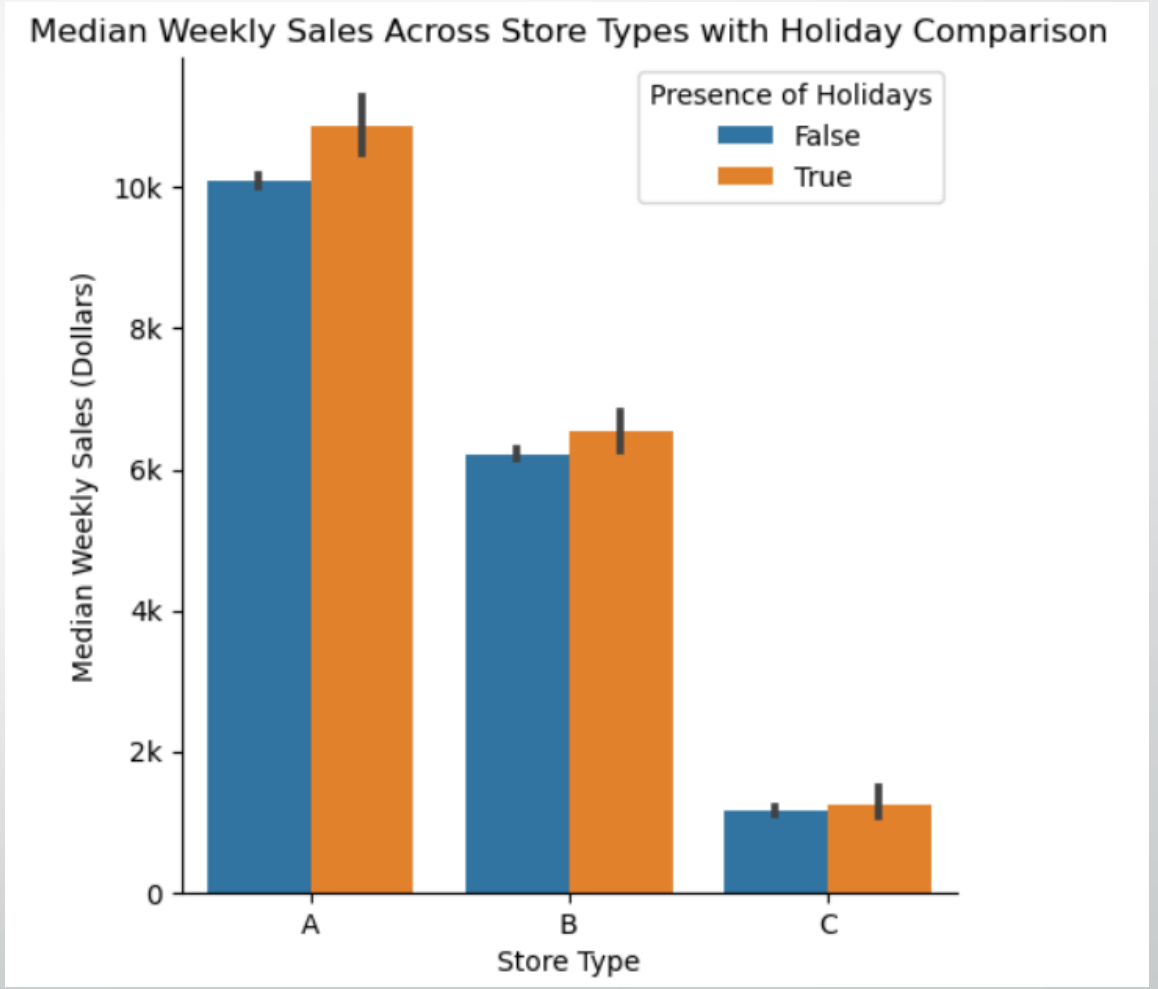


Multivariate Visualizations

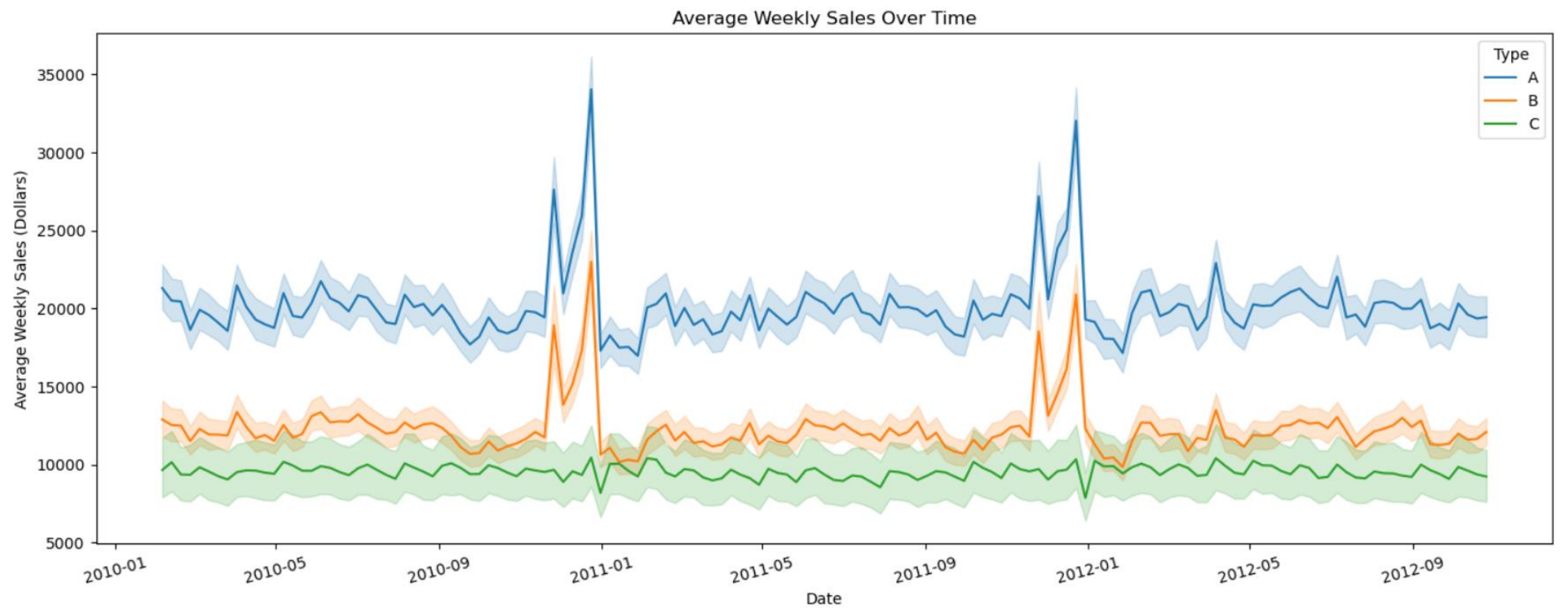
How does the average weekly sales of different types of stores differ if the week is a holiday or not?

The bar chart displays median weekly sales across store types, categorized by holiday weeks.

Store type A consistently shows the highest median sales, regardless of holidays.

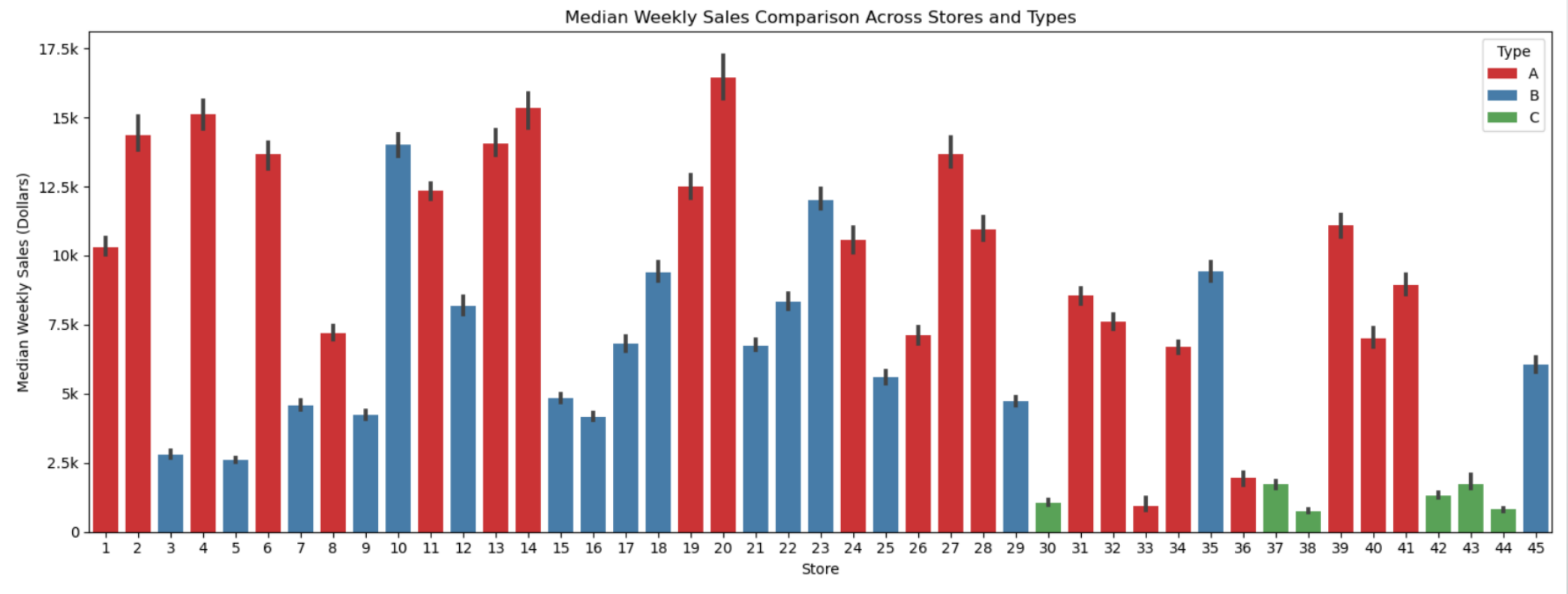


How do the weekly sales patterns of store types A, B, and C compare over the observed time period?



The line plot compares weekly sales trends among different store types. Store A consistently leads in sales, followed by B and C. Holidays show minimal impact on type C stores, possibly due to their product specialization.

What conclusions can be drawn about the performance of individual stores in relation to their types?



The analysis reveals median weekly sales patterns for store types A, B, and C. Store A consistently leads, potentially due to premium products or effective strategies. Types B and C show lower sales, influenced by varying factors like target audience and offerings.

3) Hypothesis Testing

- Holidays affects on Sales
- Types affects on Sales (A, B, C)
- MarkDowns affects on Sales

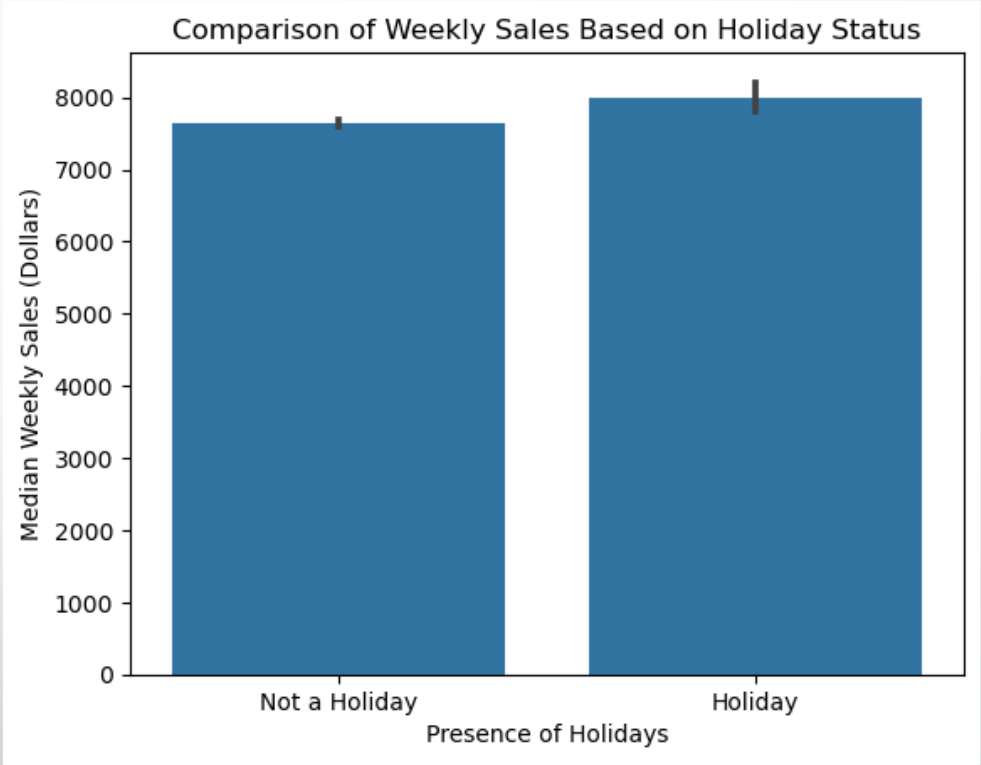
Holidays affects on Sales

Hypothesis:

- Ho: Sales in holidays = Sales in normal days
- H1: Sales in holidays > Sales in normal days

P-Value & Conclusion:

- Alpha = 0.05
- P-value = 1.4e-17 ~= 0
- Since P-Value is near Zero, we **reject the Ho** and conclude that sales in holidays are significantly higher than the sales in normal weeks.



Types affects on Sales (A, B, C)

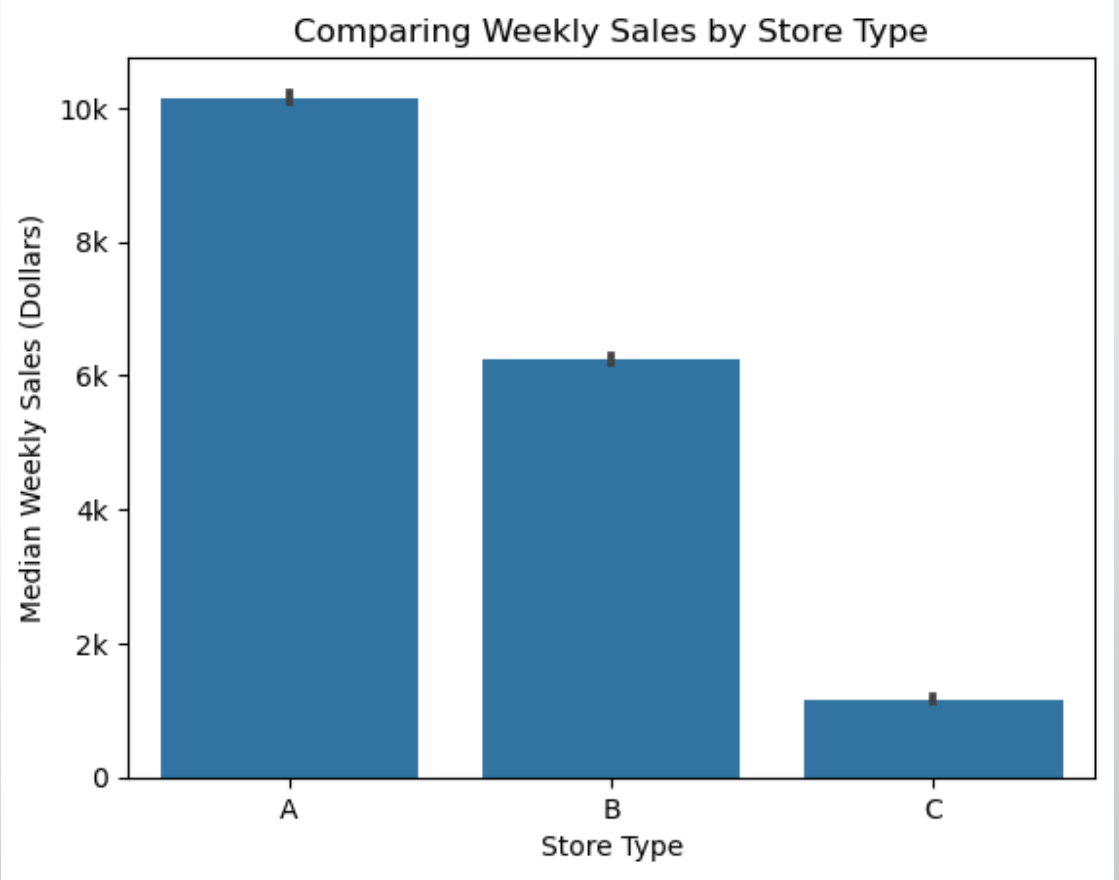
Hypothesis:

- Ho: Sales_A = Sales_B = Sales_C
- H1: Sales_A > Sales_B > Sales_C

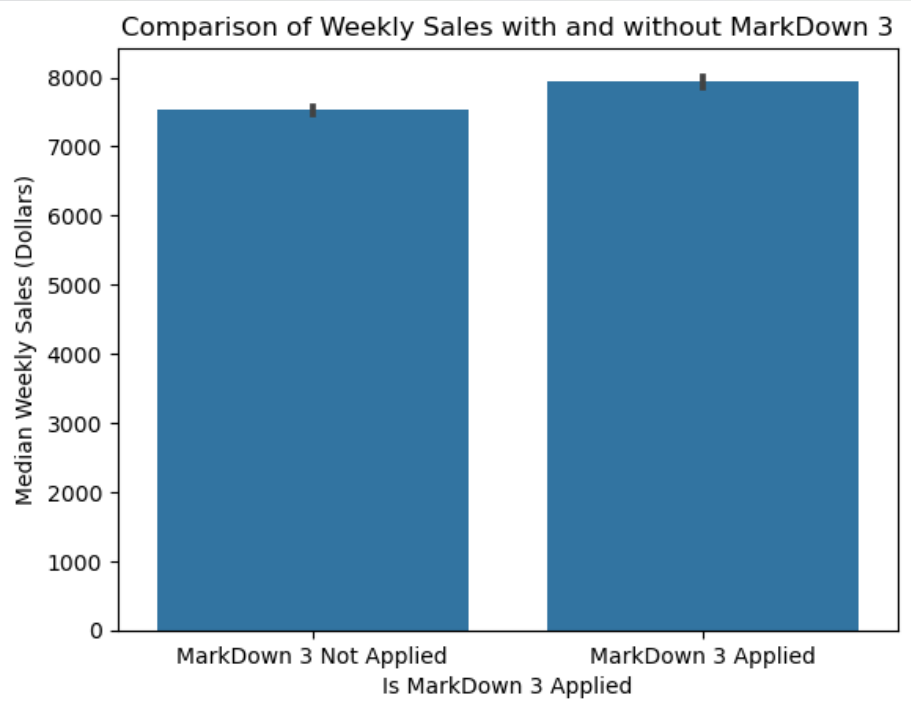
P-Value & Conclusion:

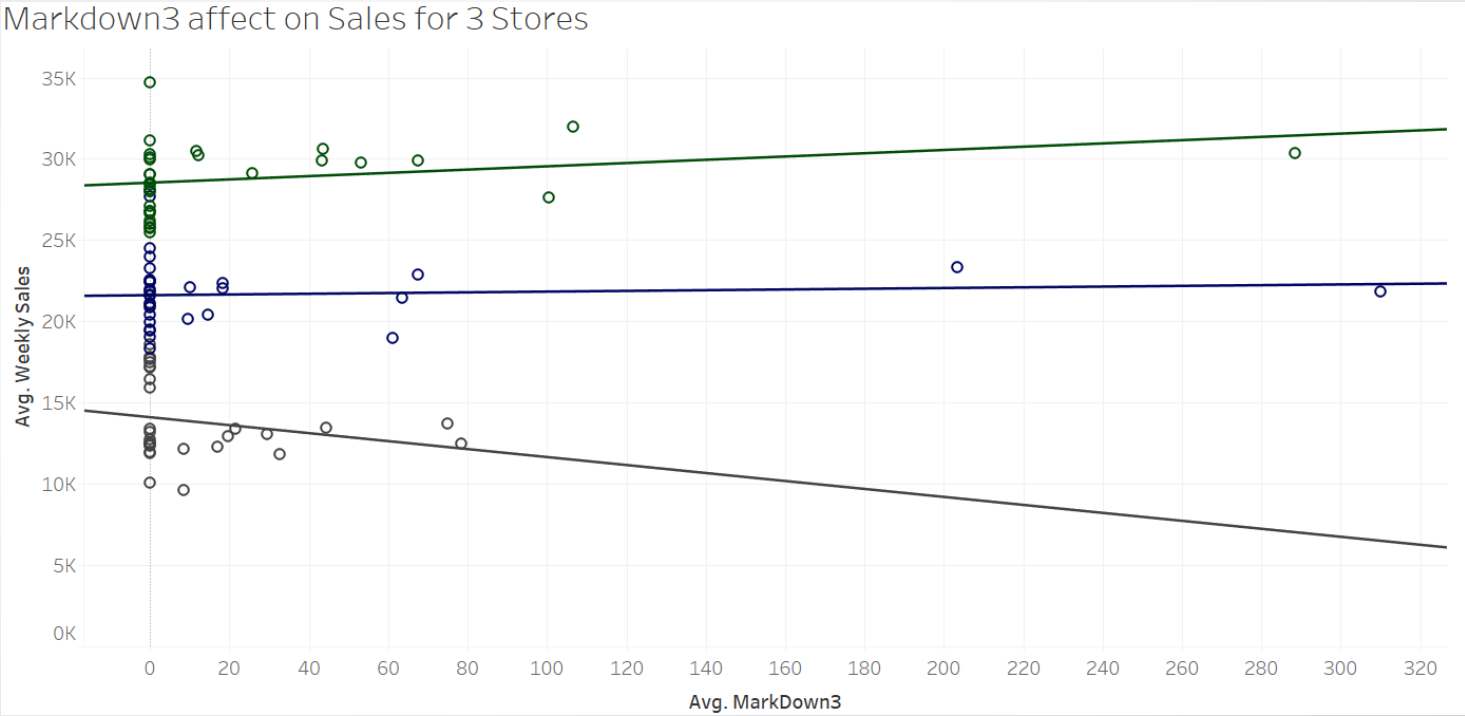
```
P_value of Sales A > Sales B =  0.0
P_value of Sales B > Sales C =  0.0
```

- **Reject Ho** and conclude that Sales is different based on Store's Type.



MarkDowns affects on Sales

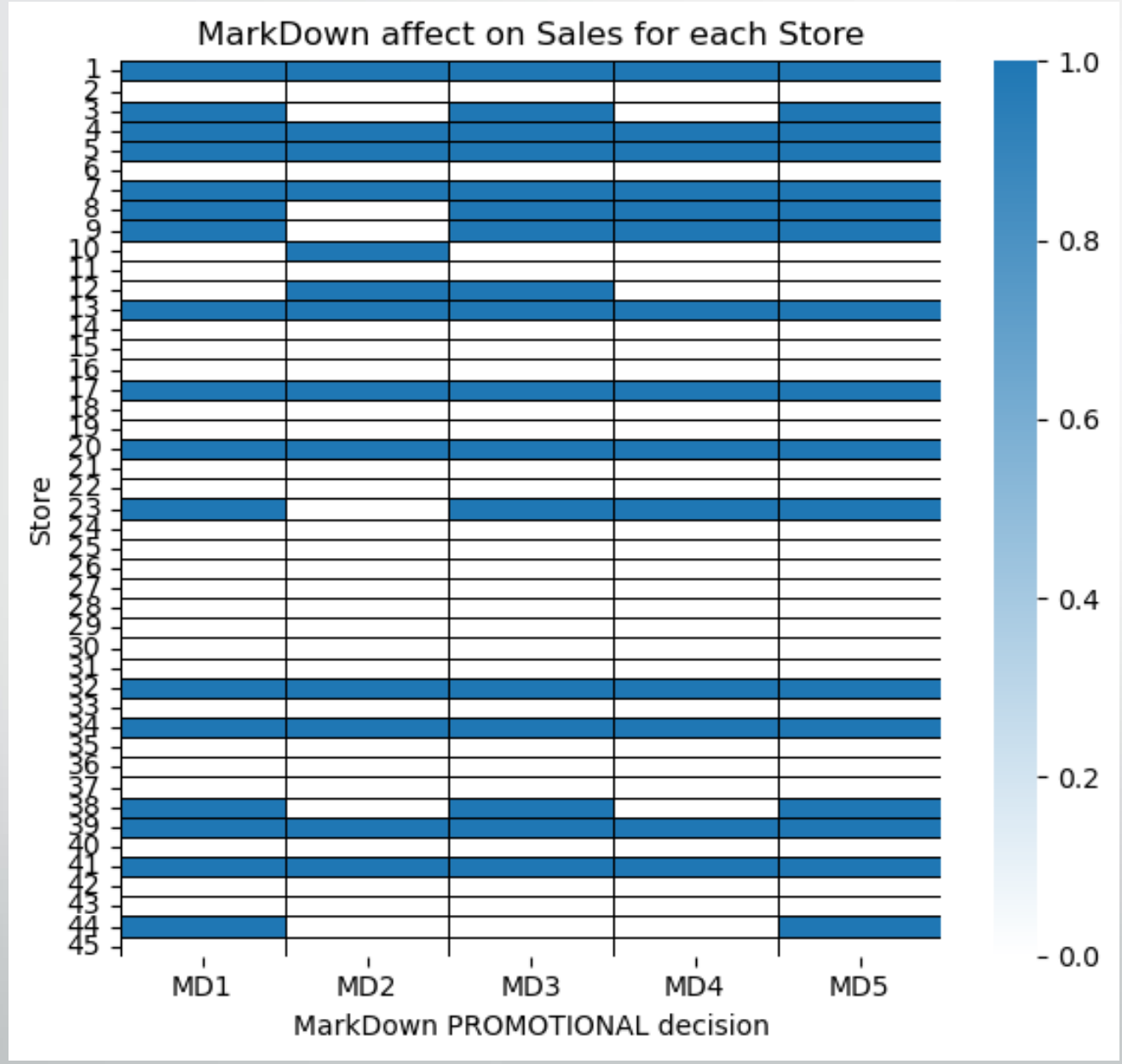




- All Stores
- Single Store Level

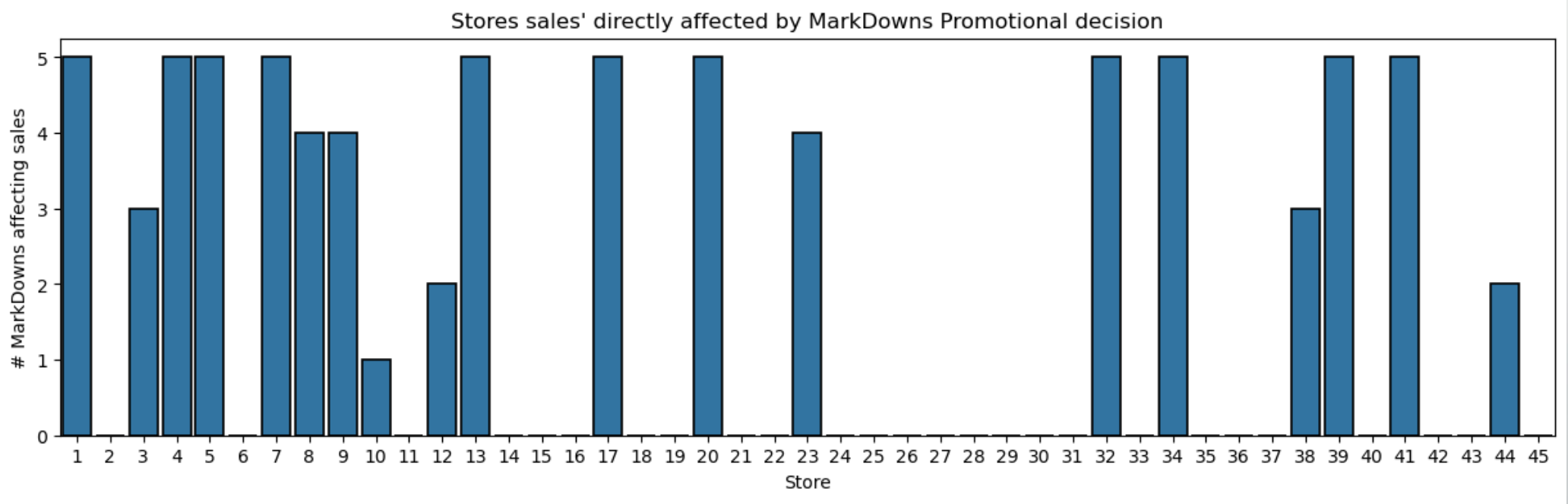
Hypothesis:

- Ho: Sales with MarkDowns = Sales without MarkDowns
- H1: Sales with MarkDowns > Sales without MarkDowns


MarkDown affect on Sales for each Store
Store
MD1
MD2
MD3
MD4
MD5
MarkDown PROMOTIONAL decision
1.0
0.8
0.6
0.4
0.2
0.0

Conclusion

- Since P-Values is less than 0.05 for some Stores we **reject Ho partially** and conclude MarkDown Promotional decisions have direct effect on certain Stores

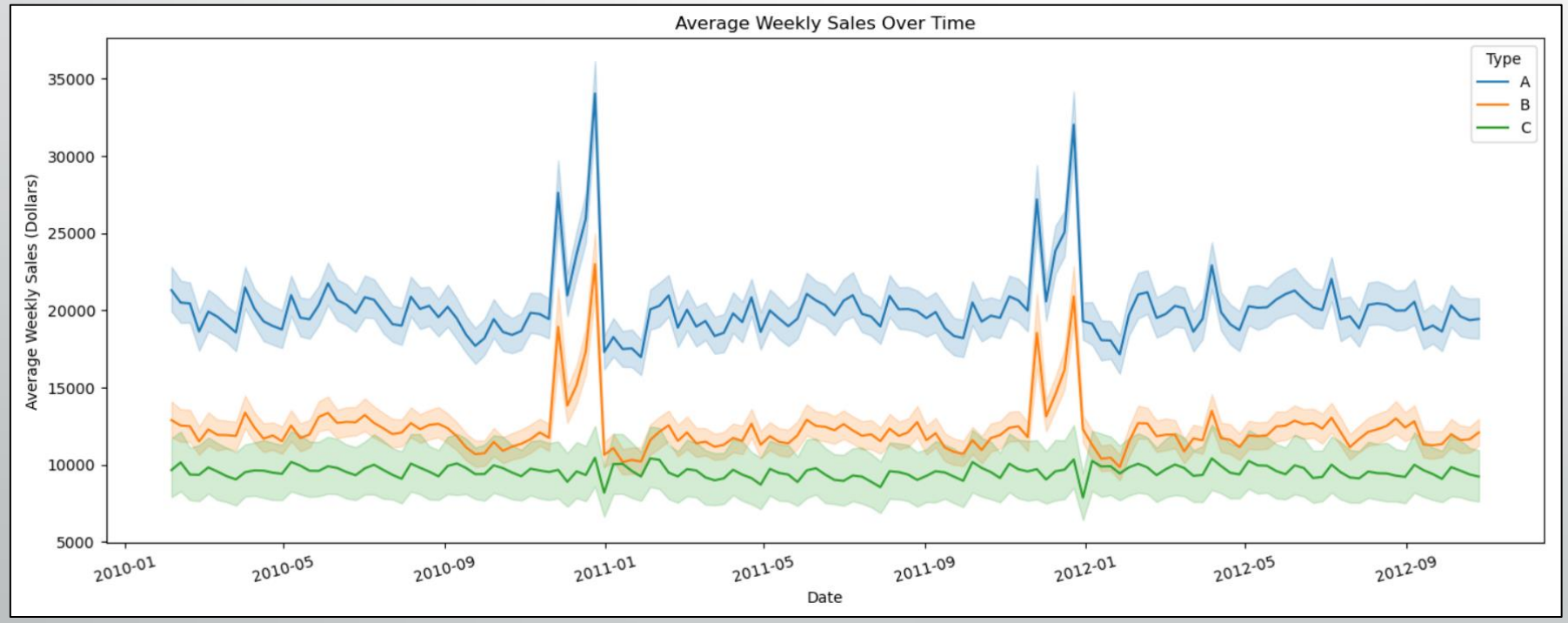


4) Feature Extraction

Its possible to extract below 2 features out of Weekly Sales Date:

- Year
- Month

We can extract Year, Month out of Week Date as both have strong relationship in determining the **Seasonality Pattern** as shown below in time series analysis of sales.



Feature Extraction Result

As expected, model prediction performance have **improved by 3%** extracting Year, Month out of Sales Week date.

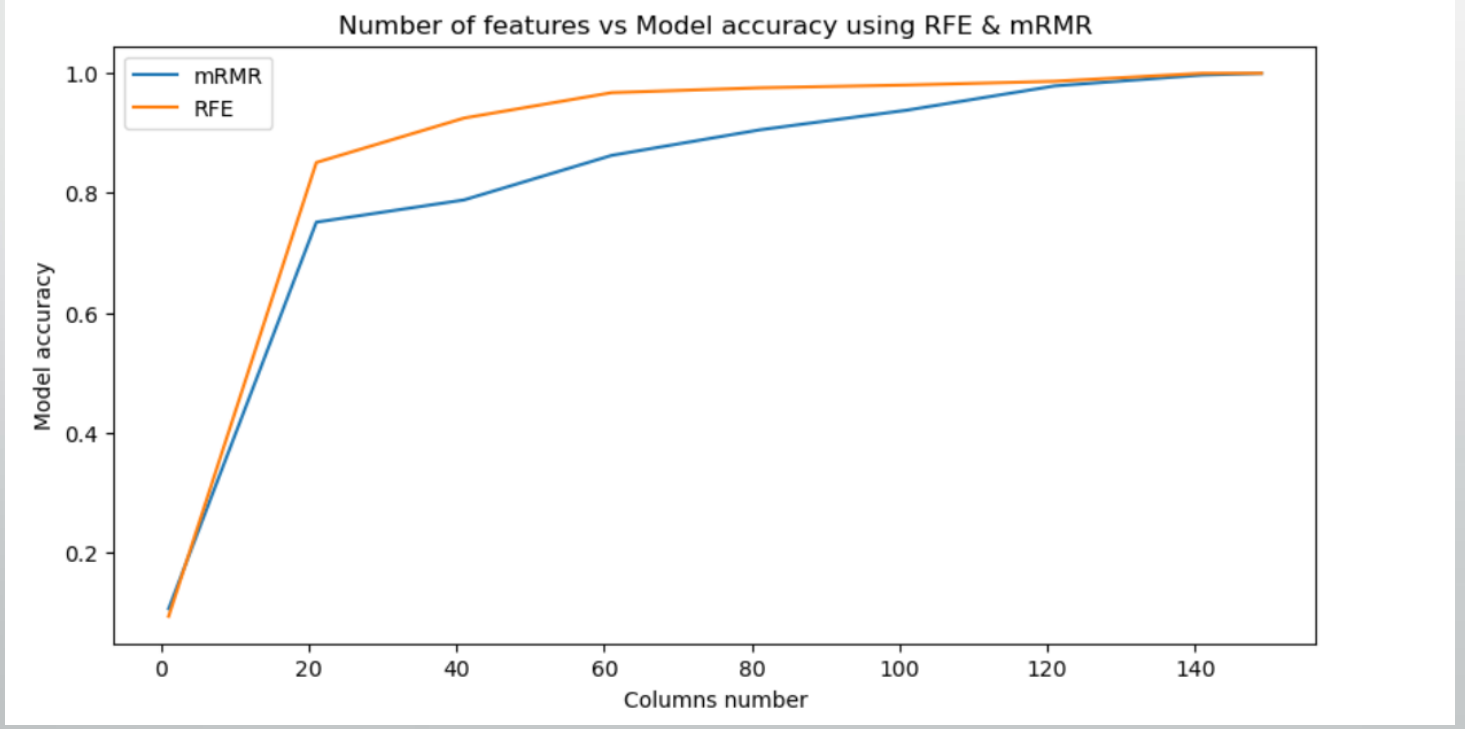
```
R2 Score without Year,Month Features: 0.91
R2 Score with Year,Month Features: 0.94
```

5) Feature Selection

We ran feature selection using **mRMR** and **RFE** models with different columns number.

The graph shows that when ever columns number increased the models accuracy will also increase!

As a conclusion, all the features in our model are relevant to the data and will include all of them in our future training model.



6) Conclusion

- Open new large branch for Stores with Type A since they have the highest Sales.
- Make use of Holidays to boost Sales by applying special events and promotions.
- Further investigation how some Stores Sales getting higher with MarkDowns promotions while other Stores not.
- Consider Economic Indicators, Climate Change, Seasonality and Store characteristics while predicting Weekly Sales for better control over the Store's resources.