Walmart

Future Sales Prediction

Group 10:

- Aditee Bhattarai
- Hemnath N
- Lindiwe Mukurazita
- Rajadurga Ganesan
- Sachet Rajbhandari
- Utsav Pradhan

SALES PREDICTION

Business Problem:

- To launch a new product in a store where the sales and demand is high

Objective:

- Identifying the correct store and time to pilot the new product to increase profitability
- Choosing the best internal and external factors which determines the weekly sales and demand

Why it is important?

- Predicting future sales for a company is one of the most important aspects of strategic planning, so that the company can make informed decisions about business planning, budgeting and risk management.

Dataset Description

- The dataset is collected from kaggle.com
- It has the historical data that covers sales from 2010 to 2012
- No of rows: 6435 | No of columns: 8

Output variables:

- Store: The store number. Ranging from 1-45.
- Date: The date of the week when each data was observed.

Input variables:

- Weekly_Sales: The sales recorded during that week.
- Holiday_Flag: value 0 represents Non-Holiday week; value 1 represents Holiday week
- Temperature: Temperature of the region during that week.
- Fuel_Price: Fuel Price in the region during that week.
- CPI: Consumer Price Index during that week.
- Unemployment: The unemployment rate during that week in the region of the store

Libraries & Packages

Libraries:

NumPy, pandas, seaborn, matplotlib

Regression packages:

From library ‘sklearn’ we used following classes to extract different regression packages for our analysis.

- Linear_model, metrics, model_selection, ensemble

Steps Taken

Data Preprocessing	Descriptive Statistics	Regression Models
Checking for & imputing missing values	Lineplot | Jointplot | Distplot	Linear Regression model
Reframing the date column	corr() function	Decision Tree regressor
Checking for outliers	Heatmap	Random Forest regressor

Data Preprocessing

Reframing date column:

- We converted the date object to ‘datetime’, using pandas.to_datetime() function
- As a next step, we reframed the date column by breaking the date into day, week, month and year categories for analysis

```
# Reframing the columns by breaking the date into weeks, month and year for analysis

df['weekday'] = df.Date.dt.weekday
df['week'] = df.Date.dt.week
df['month'] = df.Date.dt.month
df['year'] = df.Date.dt.year

df.drop(['Date'], axis=1, inplace=True)#,'month'
```

Descriptive Analysis

Number of Rows and Column

```
df.shape[0]
```

```
6435
```

```
df.shape[1]
```

```
8
```

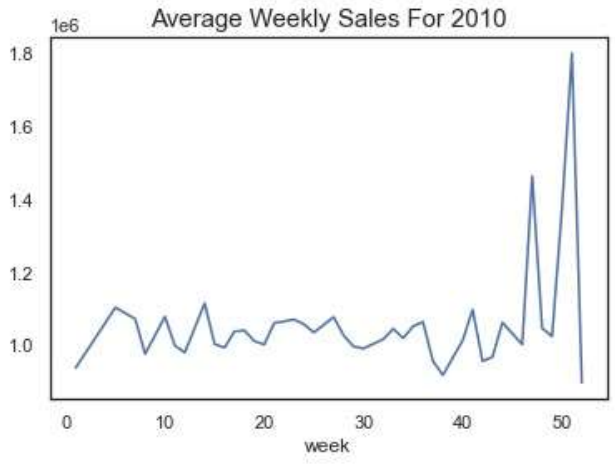
Data Summarization

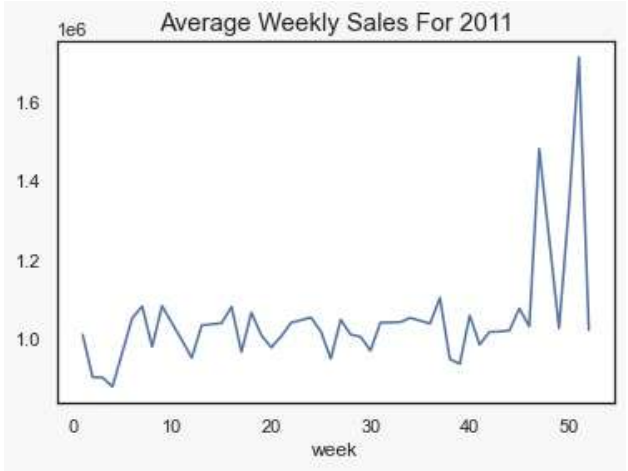
```
#getting the summary statistics of a dataframe
df.describe()
```

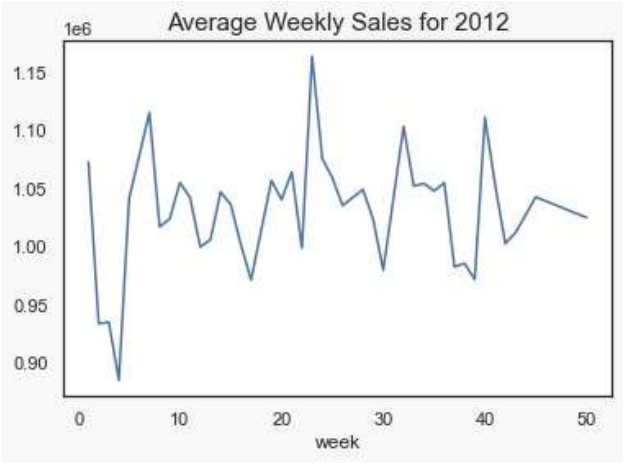
	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6435.000000	6.435000e+03	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
mean	23.000000	1.046965e+06	0.069930	60.663782	3.358607	171.578394	7.999151
std	12.988182	5.643666e+05	0.255049	18.444933	0.459020	39.356712	1.875885
min	1.000000	2.099862e+05	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	12.000000	5.533501e+05	0.000000	47.460000	2.933000	131.735000	6.891000
50%	23.000000	9.607460e+05	0.000000	62.670000	3.445000	182.616521	7.874000
75%	34.000000	1.420159e+06	0.000000	74.940000	3.735000	212.743293	8.622000
max	45.000000	3.818686e+06	1.000000	100.140000	4.468000	227.232807	14.313000

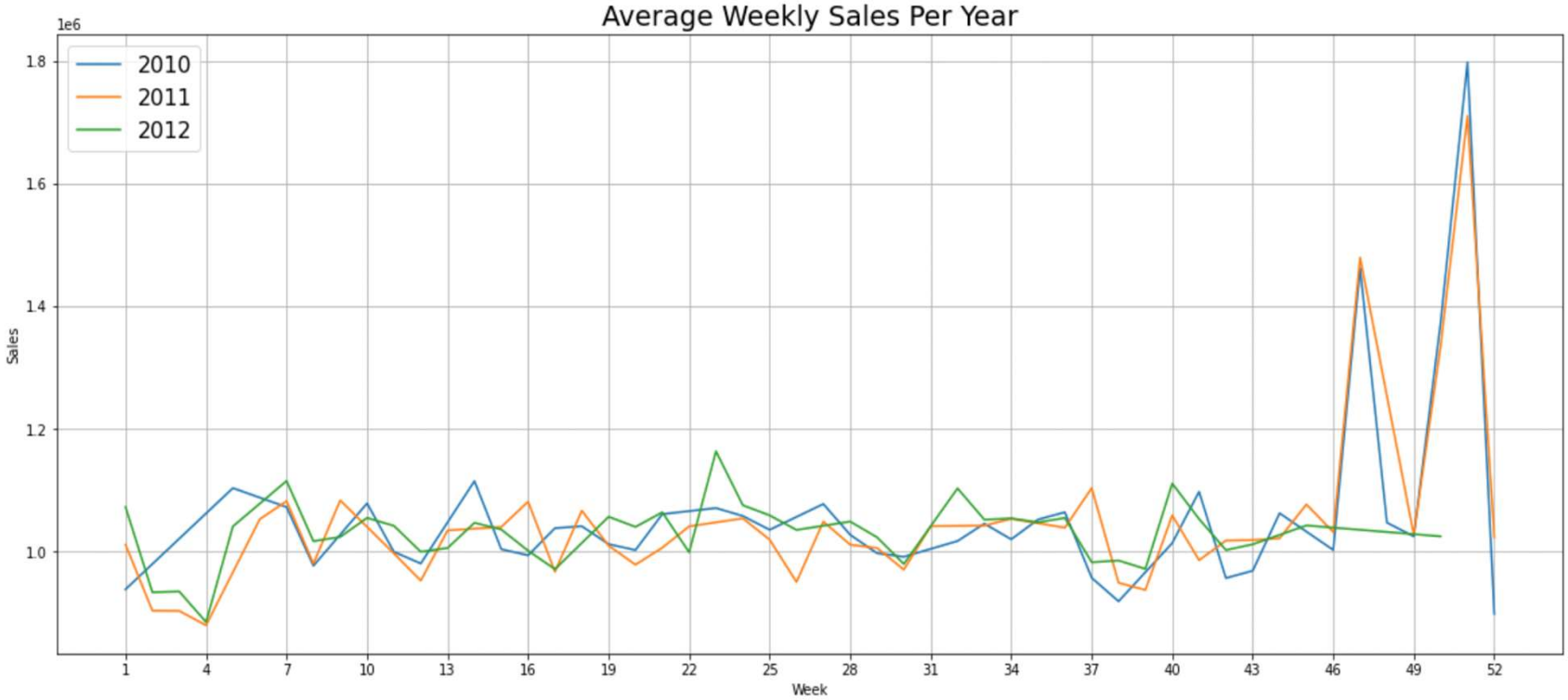
Checking for Seasonality in Weekly Sales

We used line plots to see if there was any pattern or seasonality in weekly sales over the 3 years (2010 – 2012). Our dataset, unfortunately, did not have complete yearly data for 2012.









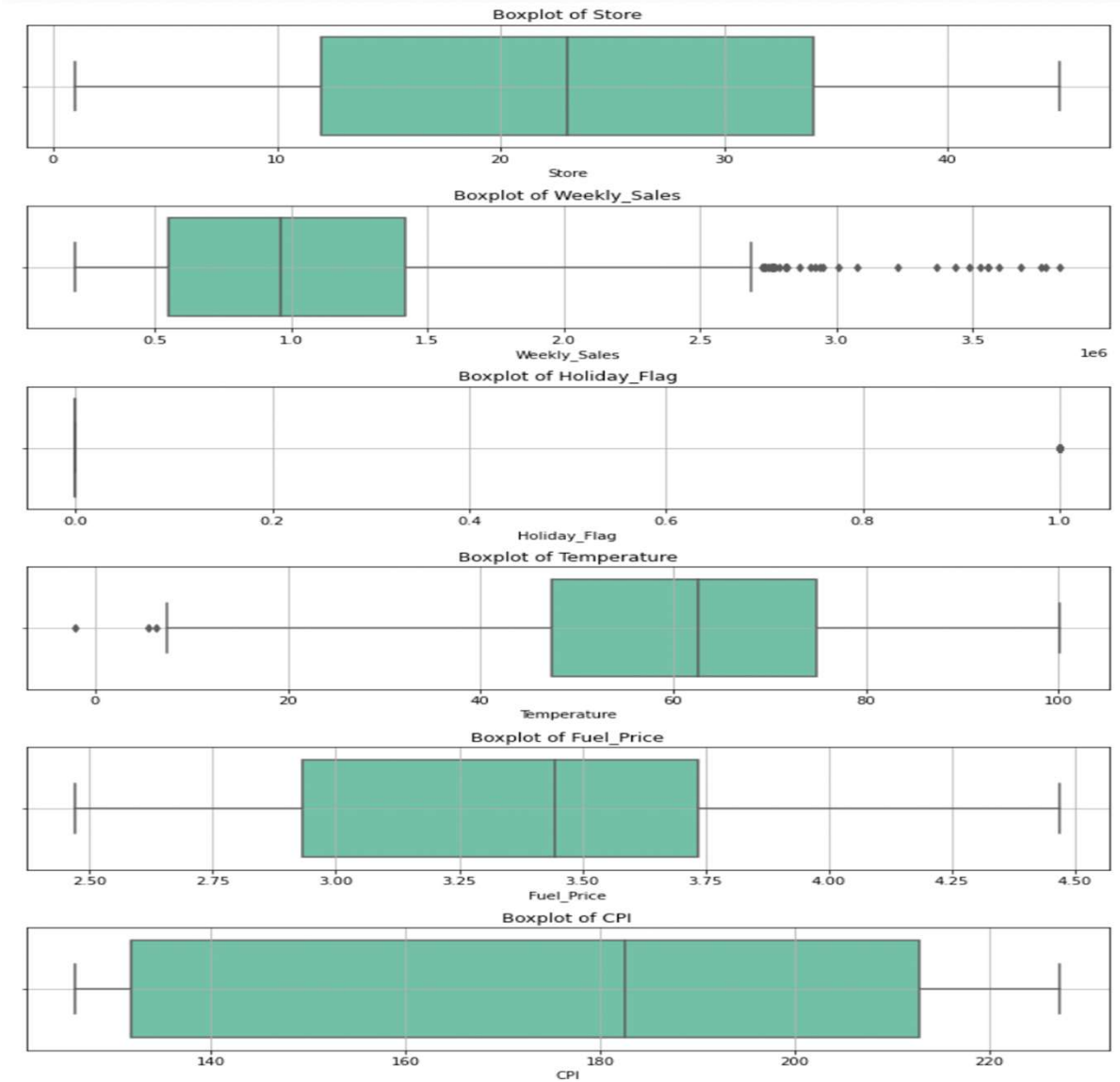
We can see a peak in **Weekly Sales** from week 46 to week 52. These weeks mark Thanksgiving & Christmas for both 2010 & 2011. Therefore, we can say that seasonality exists for Walmart. We would have to consider this factor when we later decide when to introduce our new product.

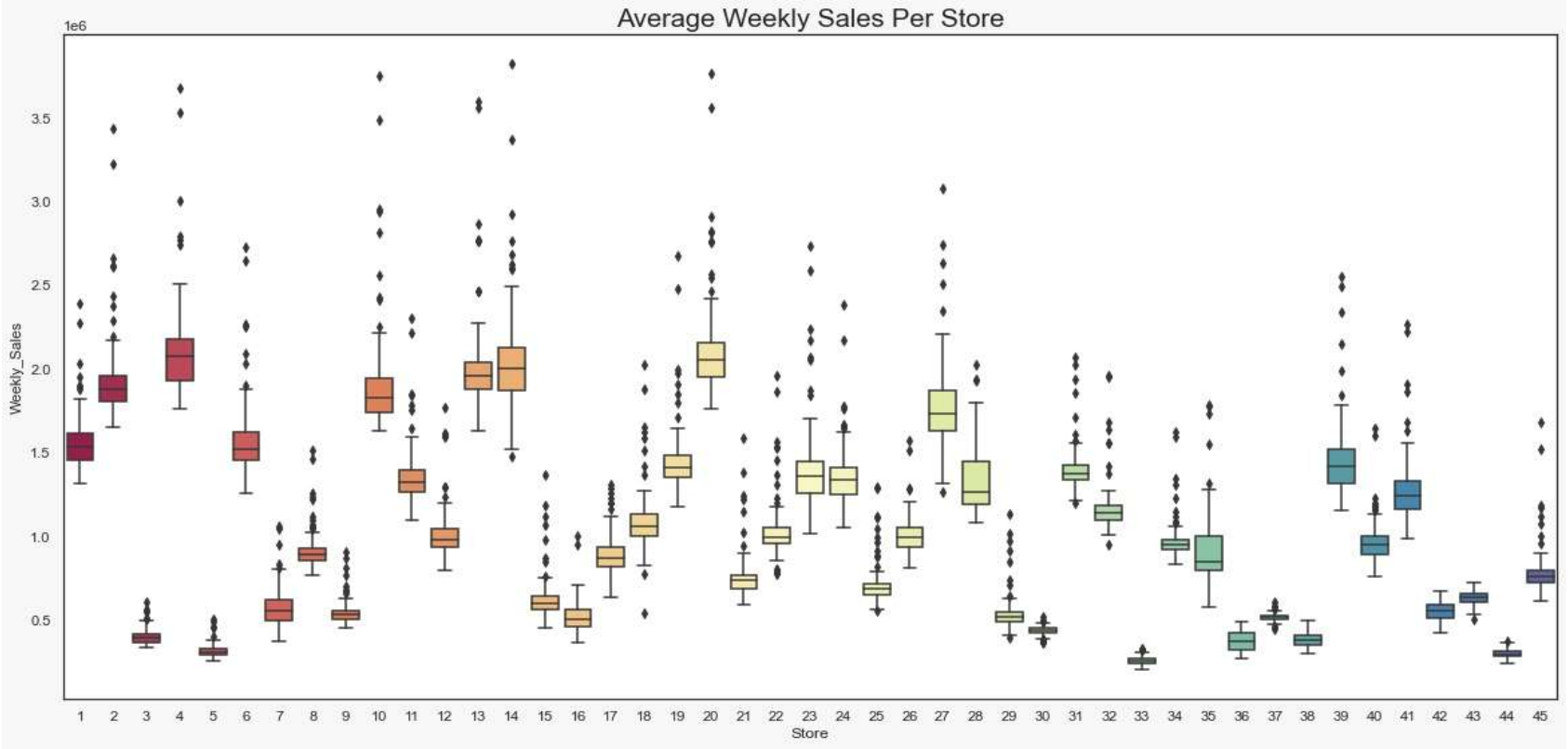
Checking for Outliers using BoxPlot

We can see that there are a lot of outliers in **Weekly Sales**. Most of these outliers can be accredited to increase in sales during the Holidays.

There are some outliers in **Temperature** as well but there are very few in number.

*(Note: **Holiday Flag** is a Boolean, thus the boxplot cannot be determined.)*





As we investigate further into **Weekly Sales**, we notice that almost all stores have outliers, which have been reflected in **Weekly Sales** boxplot. We can also clearly see that the sales is not same across all the stores. Some stores like 3, 33, 44 are underperformers while stores 4, 14 and 20 have the best sales.

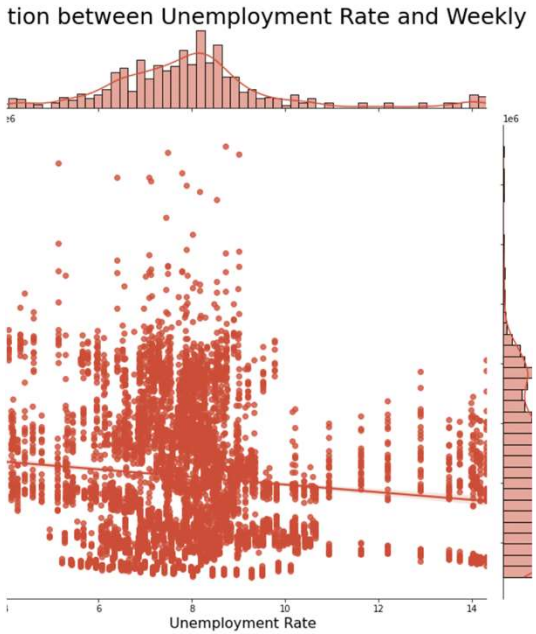
Testing for Correlation

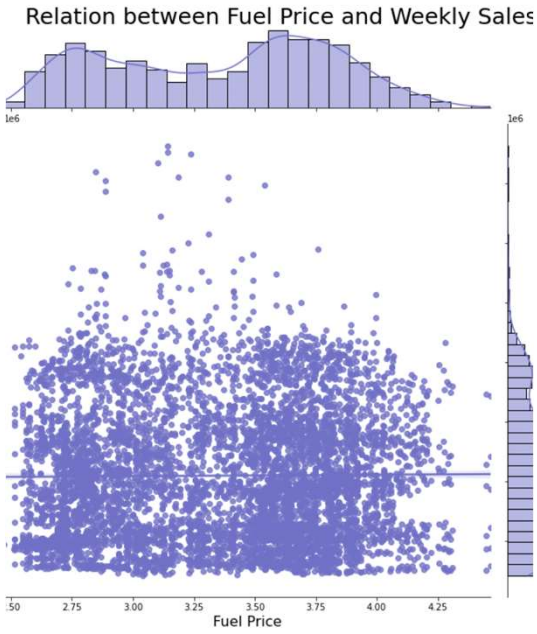
	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	weekday	week	month	year
Store	1.000000e+00	-0.335332	6.250842e-20	-0.022659	0.060023	-0.209492	0.223531	2.384098e-19	0.000000	3.071631e-19	0.000000
Weekly_Sales	-3.353320e-01	1.000000	3.689097e-02	-0.063810	0.009464	-0.072634	-0.106176	-2.104085e-02	0.066105	6.753523e-02	-0.018378
Holiday_Flag	6.250842e-20	0.036891	1.000000e+00	-0.155091	-0.078347	-0.002162	0.010960	-9.100474e-02	0.328803	3.322341e-01	-0.056783
Temperature	-2.265908e-02	-0.063810	-1.550913e-01	1.000000	0.144982	0.176888	0.101158	1.833136e-02	0.073187	6.643970e-02	0.064269
Fuel_Price	6.002295e-02	0.009464	-7.834652e-02	0.144982	1.000000	-0.170642	-0.034684	-2.651216e-02	-0.041938	-5.283174e-02	0.779470
CPI	-2.094919e-01	-0.072634	-2.162091e-03	0.176888	-0.170642	1.000000	-0.302020	-9.595877e-03	0.002783	1.478843e-03	0.074796
Unemployment	2.235313e-01	-0.106176	1.096028e-02	0.101158	-0.034684	-0.302020	1.000000	3.777320e-02	-0.006038	-2.061552e-03	-0.241813
weekday	2.384098e-19	-0.021041	-9.100474e-02	0.018331	-0.026512	-0.009596	0.037773	1.000000e+00	-0.128380	-1.387259e-01	-0.099238

Correlation Heatmap

- When we checked the data to see if there was a correlation between any of the dependent and independent variables, we found that no correlation was too strong. **Unemployment** has a negative correlation of 11% with **Sales**, which is something we will have to keep in mind while choosing a store to introduce a new product.
- **Temperature** has a negative correlation of 6% on **Sales**, which is not a very strong relationship. However, we may assume that this number is also because of the high sales we observed from the 46th week to 52nd, which is winter holiday season in the United States.

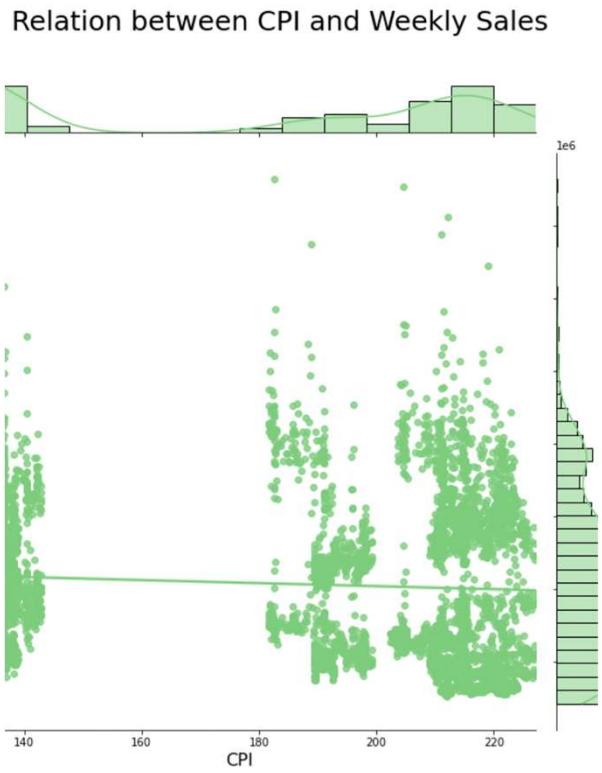
Checking for relationship between Sales and Input variables

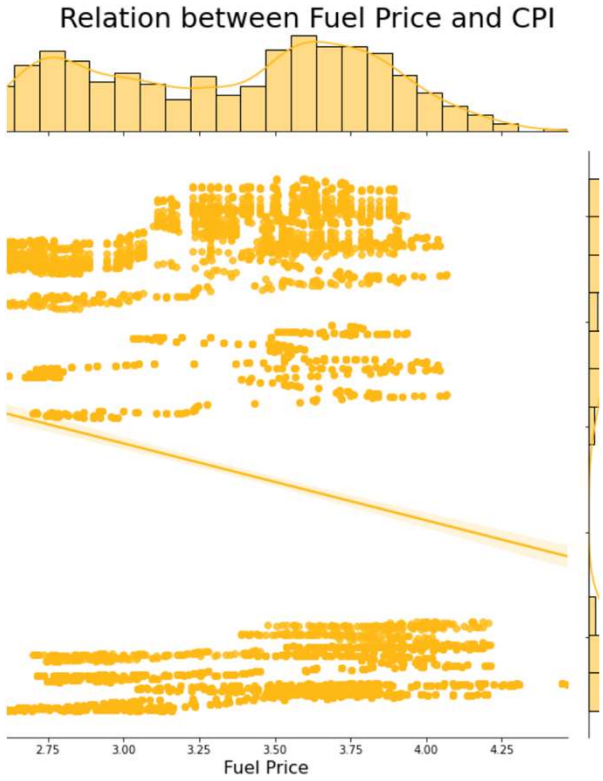




To confirm our findings from the heatmap, we plotted the **Weekly Sales** against the independent variables using joint plots with regression lines.

We can see again that **Fuel Price** does not have any effect on sales. There is a relationship between **Unemployment** and **Sales**. The store areas that have low levels of **Unemployment** have higher **Sales**.



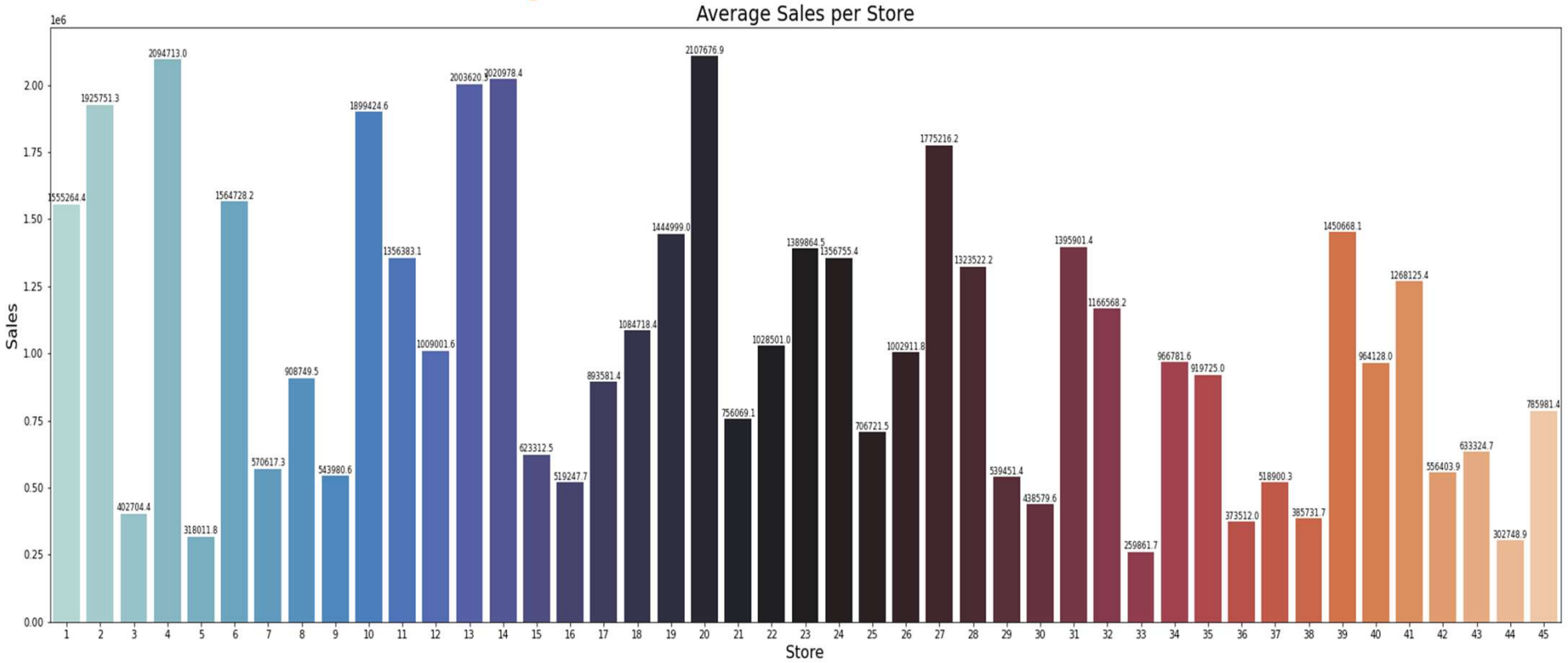


We can see that there is a very subtle relationship between **CPI** and **Sales**. This could be because product prices at Walmart are not very high, and it targets people of all income level. So even though, there might be a slight relationship, we can say that it is not significant given the product prices at Walmart.

The strongest relationship that we see in the heatmap, and the graph, is between the **Year** and the **Fuel Price**, and between the **CPI** and **Fuel Price**. However, **Fuel Price** does not have a significant effect on **Sales**.

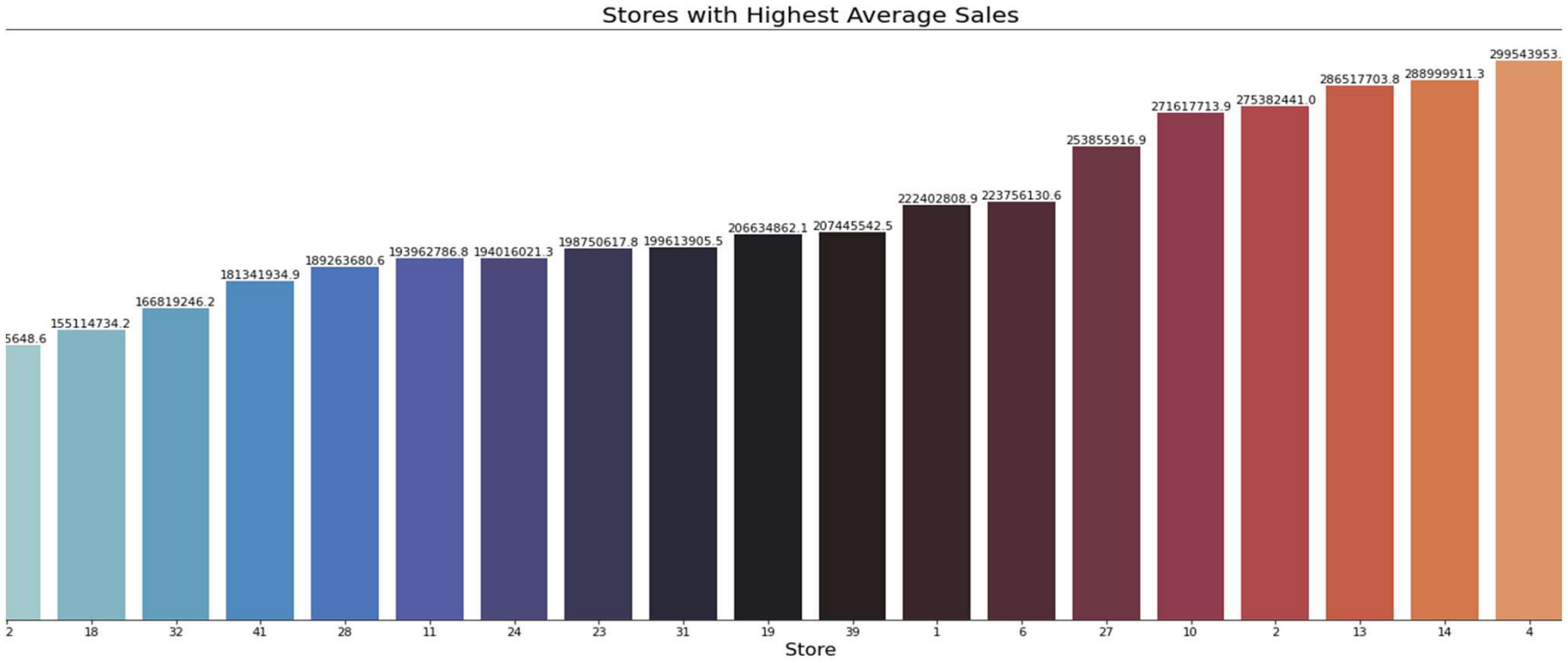
Average Weekly Sales of all Stores

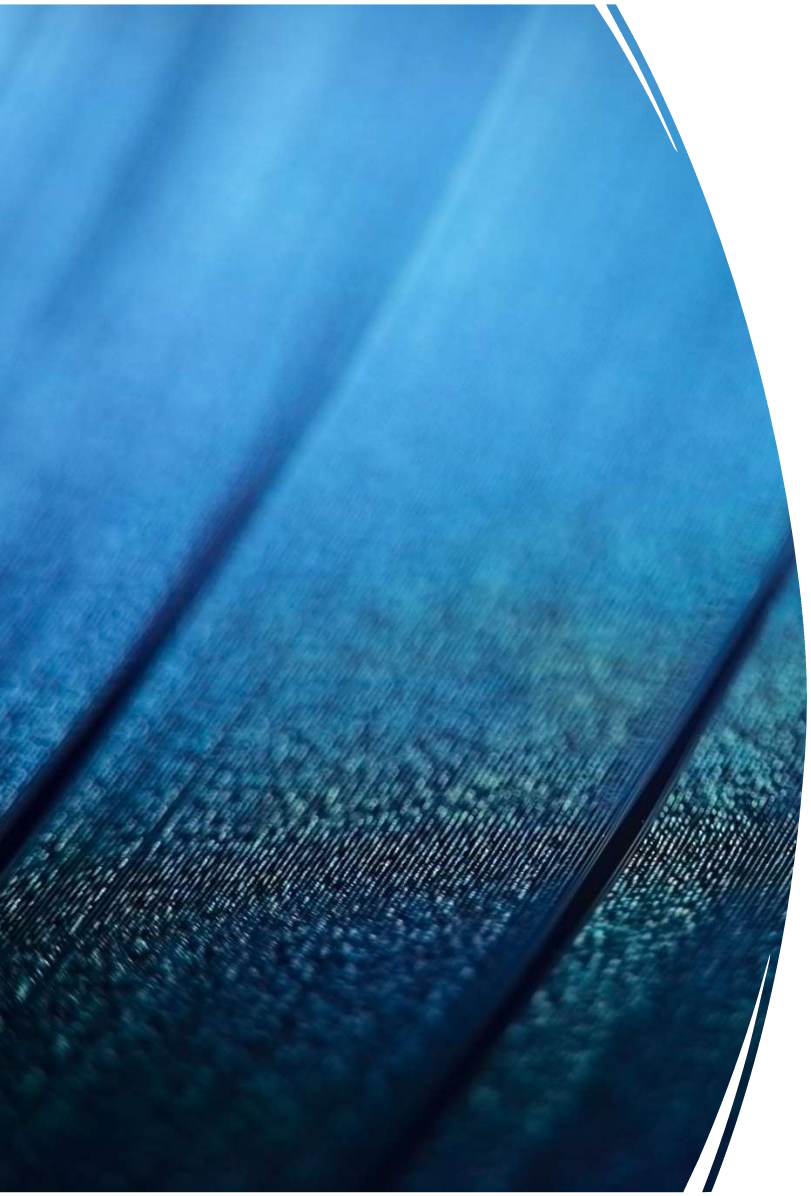
We plotted the average of **Weekly Sales** of all the stores from 2010 to 2012 to identify the stores with the highest sales.



Average Weekly Sales of all Stores

We picked the top 20 stores with highest **Sales** and plotted them in the bar graph. We found that the top 5 performing stores (based on sales) were Store: 4, 14, 13, 2 and 10. We would like to focus our efforts on these stores when we introduce a new product.





Top 5 Stores with highest average Weekly Sales

Store 20	$301,397,800
Store 4	$299,544,000
Store 14	$288,999,900
Store 13	$286,517,700
Store 2	$275,382,400

Prediction Model

We ran 3 regression models for test of fit (Linear Regression, Decision Tree Regressor, Random Forest Regressor) on our train data set and executed it in our test data set.

Based on the test score, we decided to go with Random forest regressor as it had the highest fit of 0.946745.

	Train RMSE	Test RMSE	Training Score
Linear Regression	144292.189330	153796.506068	0.934094
Decision Tree Regressor	0.000000	180089.006541	1.000000
Random Forest Regressor	51498.060776	131413.091509	0.991605

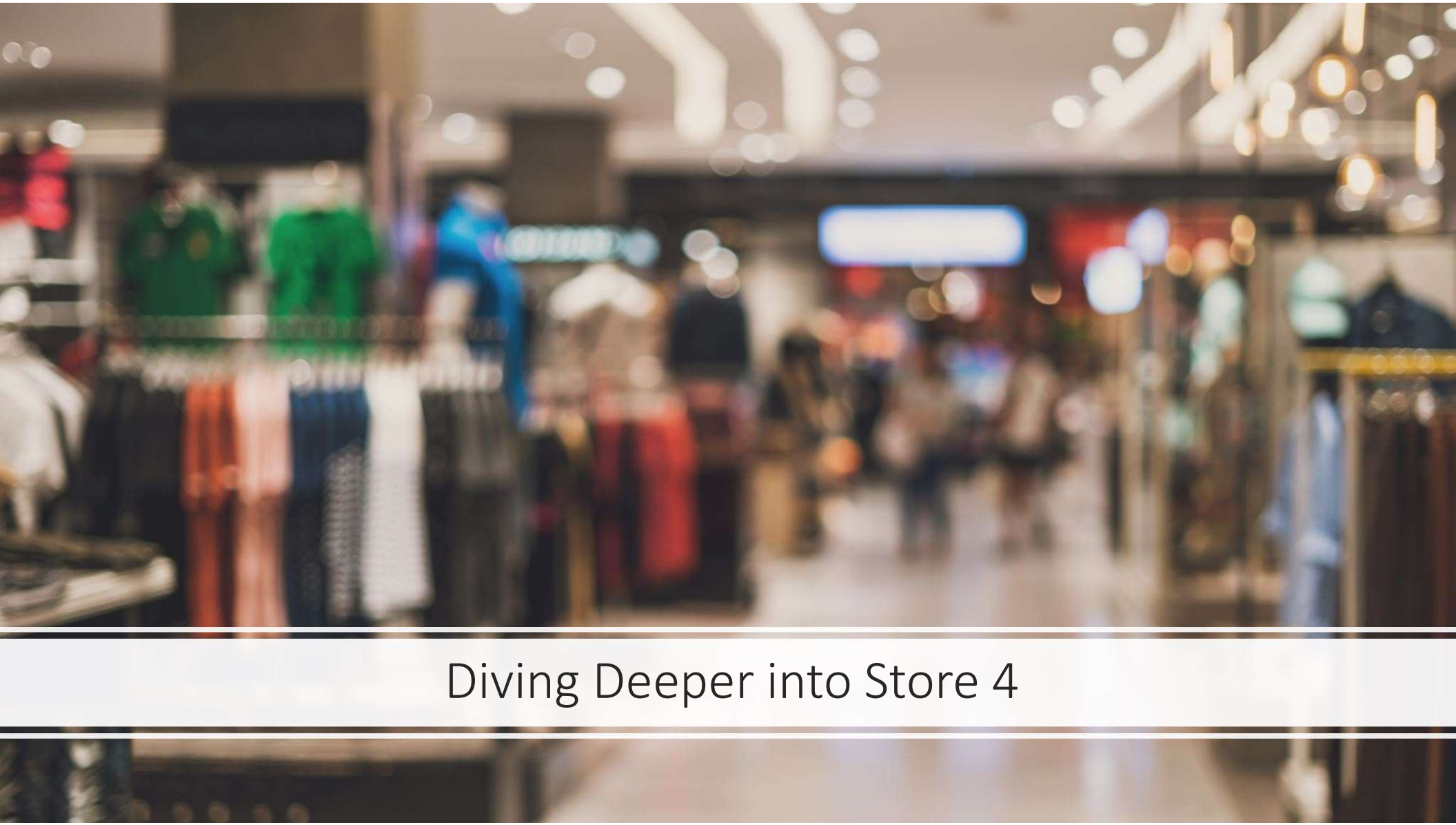
	Test Score
Linear Regression	0.927058
Decision Tree Regressor	0.899987
Random Forest Regressor	0.946745

Top 5 Most Important Stores	
Store_4	0.083731
Store_20	0.082946
Store_13	0.080040
Store_14	0.078794
Store_2	0.068834

A Random Forest is a technique capable of performing both regression and classification tasks. It combines multiple decision trees in determining the final output rather than relying on individual decision trees. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model.

Based on the Test Scores, we performed Random Forest Regression with the motive to find the best or most important store is the on all the factors considered from our dataset.

From the regression model, we see that **Store 4** is the most important store.



Diving Deeper into Store 4

- We sorted out Top 5 **Stores** based on **Total Sales** with highest sales at the first.
- Store 4 ranks 2nd in terms of Highest sales.
- We also sorted out **Stores** based on **CPI**, **Unemployment** and **Fuel** with least in the first.
- Store 4 ranks 3rd in terms of lowest **CPI** and **Fuel** prices, indicating that the location is relatively cheaper to live and travel in.
- Store 4 ranks 3rd in terms of least **Unemployment** rate as well, which indicates that more people in that location are employed and can potentially purchase our new product.
- Based on Random Forest Regression model and our previous analyses, we can conclude that Store 4 is the most ideal store for us to introduce our new product.

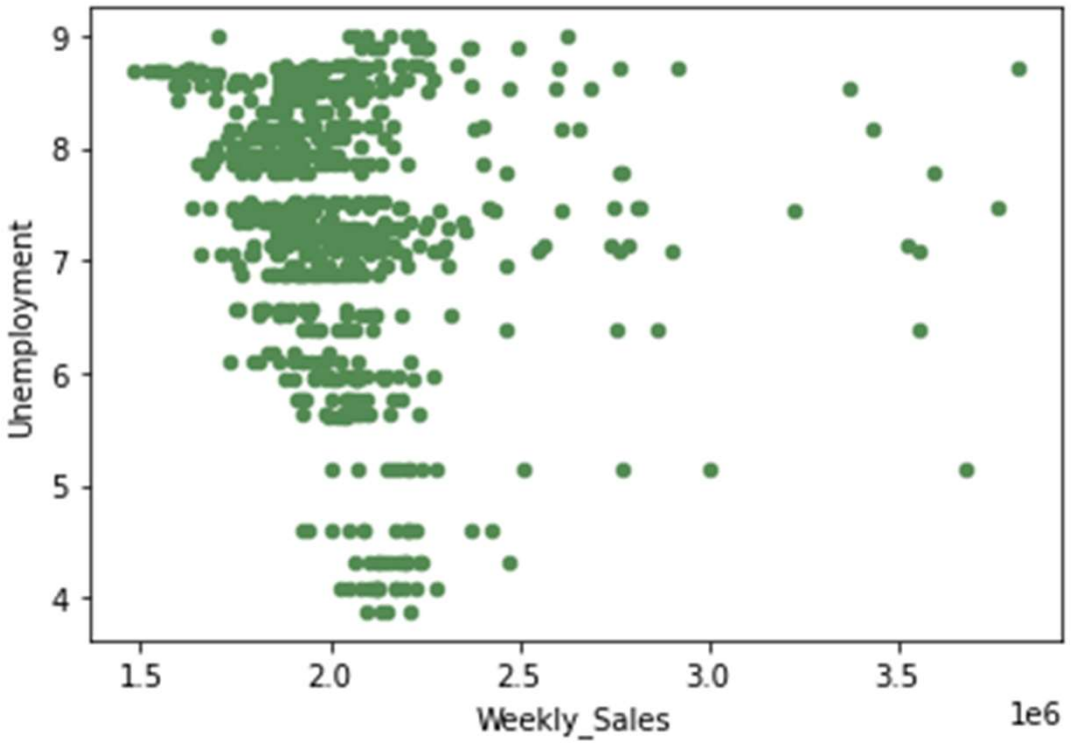
Store	Total sales
20	301,397,800
4	299,544,000
14	288,999,900
13	286,517,700
2	275,382,400

Store	CPI
28	128.679669
42	128.679669
4	128.679669
38	128.679669
34	128.679669

Store	Unemployment
23	4.796014
40	4.796014
4	5.964692
8	6.091846
9	6.099881

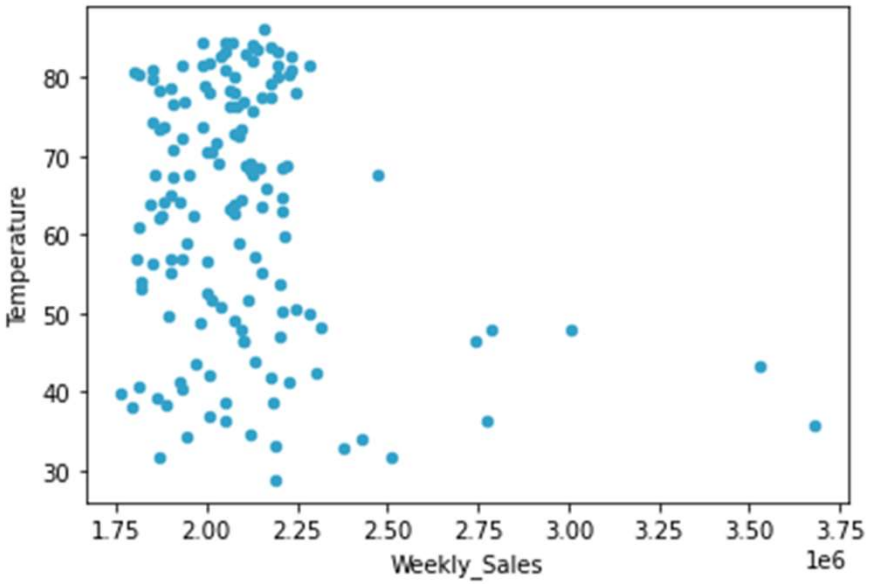
Store	Fuel
36	3.204203
34	3.216972
4	3.216972
1	3.219699
21	3.219699

Relationship between variables in Store 4

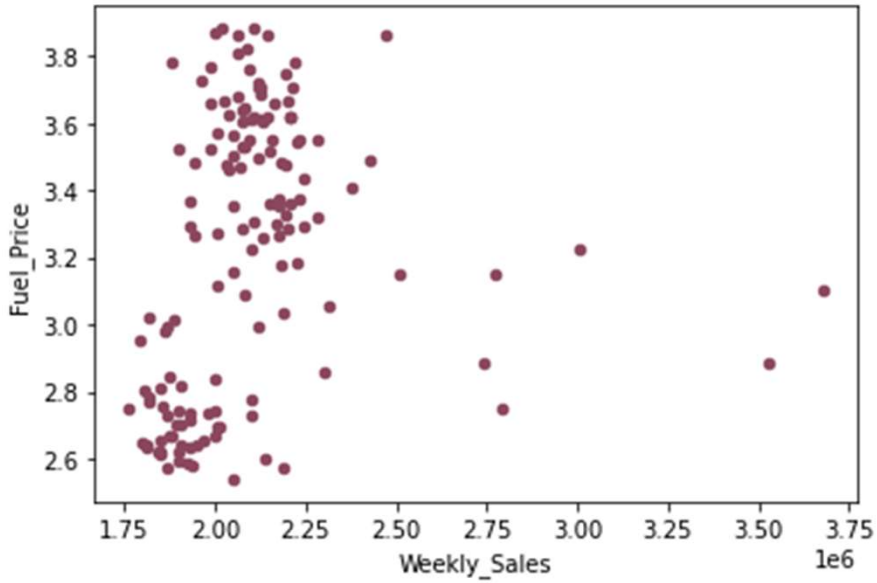


After running Random Forest Regressor, initially on all the Stores, and later in Store 4, we wanted to look deeper into the factors that were affecting the **Sales** in this store.

Here we can discern that **Unemployment** seems to have some impact on Sales, such that Higher unemployment leads to lower weekly sales.



Weekly Sales in Store 4 are marginally higher when the **Temperature** is warmer.



Increase in **Fuel** prices does not seem to affect the **Weekly Sales** in Store 4.

Recommendation & Insights

- Store 4 is the most ideal store
  - One of the Highest Sales Value
  - Low Unemployment, CPI, and Fuel prices
- Introduce the product between week 46 & 52
- Unemployment Rate, among other factors, plays a relatively significant role in determining sales.