MDA 620

CAPSTONE PROJECT

“WALMART SALES PREDICTION”

Introduction

In our project we decided to go over Walmart Store Sales that cover the period of time from 2010/02/05 to 2012/11/01 and that take into account 45 different stores located in different Regions.

Inside the data set we will find 8 different variables, 2 categorical and 6 numeric variables. Each of these variables is more or less correlated with the Weekly Sales made by each Walmart store, we would like to see which variable has more impact on the weekly sales and if it is positive or negative.

Once we have done the analysis of our variables we will look more into the details about the predictions of sales, we will build a model that would be able to predict the future sales of Walmart based on the previous sales and other economic variables.

Our Data set there will be divided in 8 different columns and 6435 rows.

Column Explanation

Store	The store Number
Date	The store Number
Weekly_Sales	Sales for the given store

Holiday_Flag	Whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
Temperature	Temperature on the day of sale
Fuel_Price	Cost of fuel in the region
CPI	Prevailing consumer price index
Unemployment	Prevailing unemployment rate

The Holiday_Flag column will take into account the following Holiday Events dates:

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

Problem Analysis

Walmart Inc. is a US multinational, owner of the Walmart retail chain of the same name, founded by Sam Walton in 1962. It is the largest chain in the world in the large-scale distribution channel with, as of December 2021, 11,847 stores and clubs in 27 countries.

In our Project we would like to predict the sales and demand of Walmart products. As you can already see in the columns we will be using, there are some specific events and holidays that might influence the sales.

Before and during these “Special weeks” Walmart used to put in place specific Markdowns, in order to attract more customers and make more sales.

This is why during Christmas, Thanksgiving, Labour Day or the Superbowl, the weight of our weeks will be weighted higher in our evaluation than “Normal weeks”.

Other factors that are to be taken into account are the economic conditions each Walmart is under, as we said previously in the data set we will be using data of 45 stores of Walmart located in different regions, this therefore means that we will have different conditions in each region and each year.

Sometimes, due to higher demand, Walmart can have some problems, such as run out of stock of products, especially during the holiday seasons or when the economic conditions are more favorable for customers. This is why we thought that having a model that predicts the sales demand taking into account economic condition, Unemployment Index, CPI and holidays would be the solution to the problem.

Objective of the project

- Filter and understand the dataset
- Create a prediction model on sales
- Evaluate the different models and choose the best one for prediction

Data Exploration

For starting we uploaded the dataset on github and created a direct link for it, in that way we can use it for our data set analysis and everyone can have access to it.

```
import pandas as pd
import matplotlib.pyplot as plt
walmart =
pd.read_csv("https://raw.githubusercontent.com/elinor00/walmart/main/Walmart.csv")
```

We will have 6435 rows and 8 different columns, and before starting we wanted to make sure that we won't have too many NAs. When building a prediction model having missing values could cause a misinterpretation of the data, and lead to a not very accurate prediction.

To avoid this, we have decided that in case we have too many NAs in a column we would consider the elimination of the column.

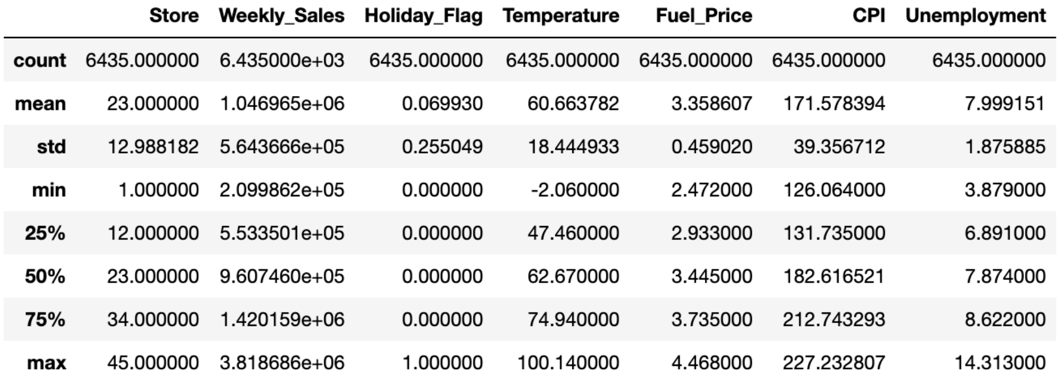
```
walmart.isna().sum()
```

Once we ran the code we saw that fortunately our dataset doesn’t have any NAs, so we can go ahead with the exploration of the data into more details.

We thought it could also be useful for us to have a description and a summary of all the columns.

In the visualization below we can see the count, mean, standard deviation, minimum and maximum, 25%, 50% and 75% percentile.

walmart.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

Data Visualization

In this section we are going to create different data visualizations in that way we can better understand the data.

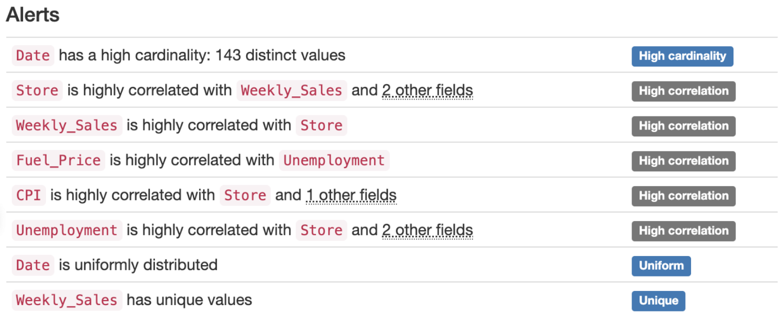
pip install pandas-profiling

profile = ProfileReport(walmart)

profile

We decided to import pandas-profiling in order to have an immediate overview of our data set and a small summary that we already discussed in the previous section.

The most interesting side in our opinion is that you can see how variables are correlated to each other, this means how one variable can influence the other.



Alerts

`Date` has a high cardinality: 143 distinct values	High cardinality
`Store` is highly correlated with `Weekly_Sales` and 2 other fields	High correlation
`Weekly_Sales` is highly correlated with `Store`	High correlation
`Fuel_Price` is highly correlated with `Unemployment`	High correlation
`CPI` is highly correlated with `Store` and 1 other fields	High correlation
`Unemployment` is highly correlated with `Store` and 2 other fields	High correlation
`Date` is uniformly distributed	Uniform
`Weekly_Sales` has unique values	Unique

The second section of the profiling is an analysis and visualization of each variable we have in our data set.

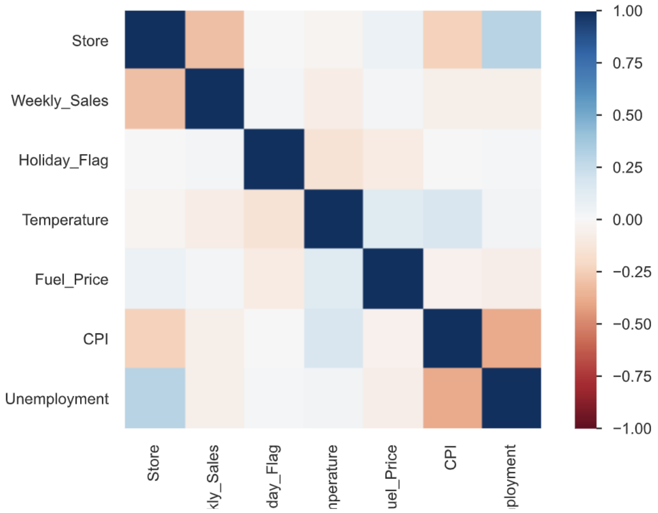
We can see a description of it, a summary (mean, 25%,50%,75%) and a graph that will tell us how the frequencies of the values are distributed.

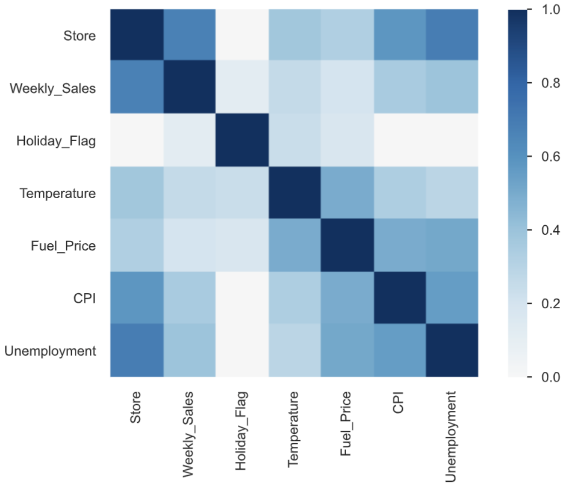
And in conclusion, the most interesting thing that we can see inside the profile report are the interactions and correlations between variables.

We decided to get more into deep about the Spearman’s rank correlation coefficient.

It is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's r.

Its value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.





The Phik (φk) is a new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution.

From the graph we can see that in the zones where the color is darker it means we have a high correlation between the two values, for example CPI and Unemployment, or unemployment and stores or weekly sales with the store.

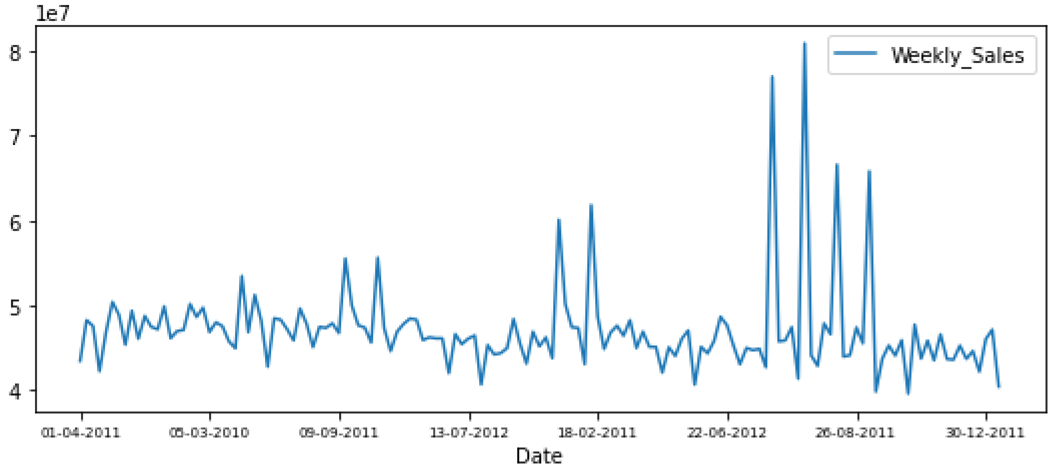
To get more deep into it we decided to create some other visualizations that will mostly tell us how the weekly sales variate based on some of the economic indicators we have and the holidays week.

1. Date and Weekly Sales - Plot Line

Starting, we thought it would be interesting to have a general view of the Sales, how they are distributed and how the holidays affect them.

From this graph we can see the different sales over the years we have into consideration: 2010-2011-2012.

We can see that the most sales were made over the year between 2011 and 2012.



2. Sales During Special Holiday Weeks and Normal Weeks - Pie Chart

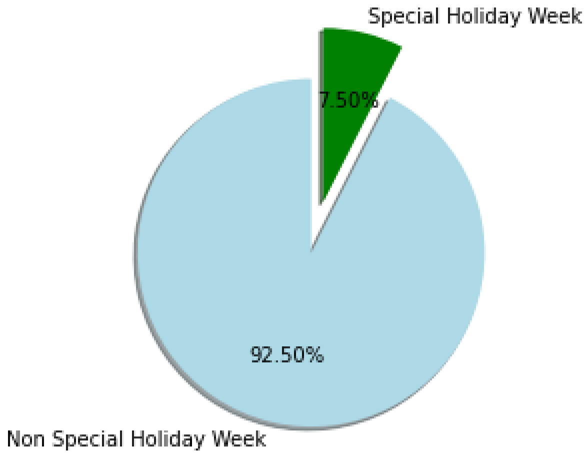
Now let’s analyze the holidays.

Analyzing the graph below we can see that 7.50% of the sales of Walmart are done during Special Holiday Weeks.

This data compared to the total amount of sales is not to undervalue. Almost 10% of the annual sales are made during holidays.

Looking more deeply into it, as we said at the beginning of our project the deals that Walmart puts in place are well received by the customers, and they have a positive impact over the total annual sales.

This gives us a better understanding and a more defined view of how this variable affects the store.



3. Weekly sales and Temperature - Histogram

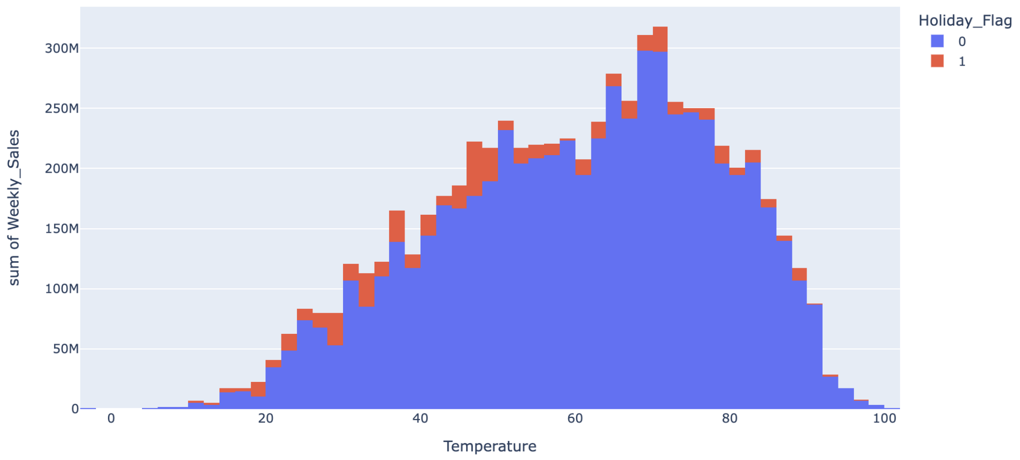
In our dataset there are different conditions that might affect our weekly sales, one of them is the Temperature.

In this graph we can see how the temperature affects the sales over the year.

We also used to fill up our graph with the holiday_flag column.

The holiday flag column will tell us whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week.

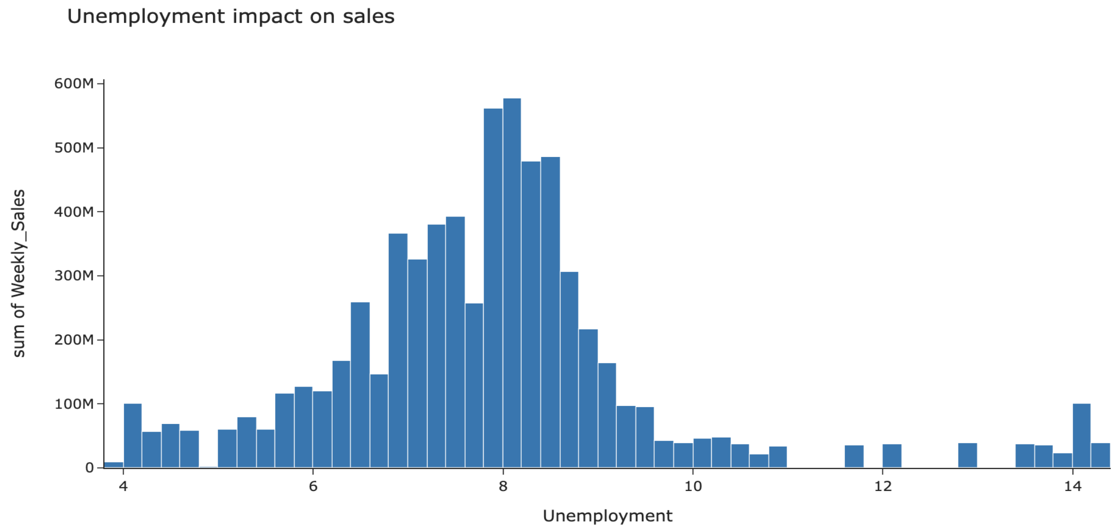
We can see how when the temperature is higher the sales go up as well.

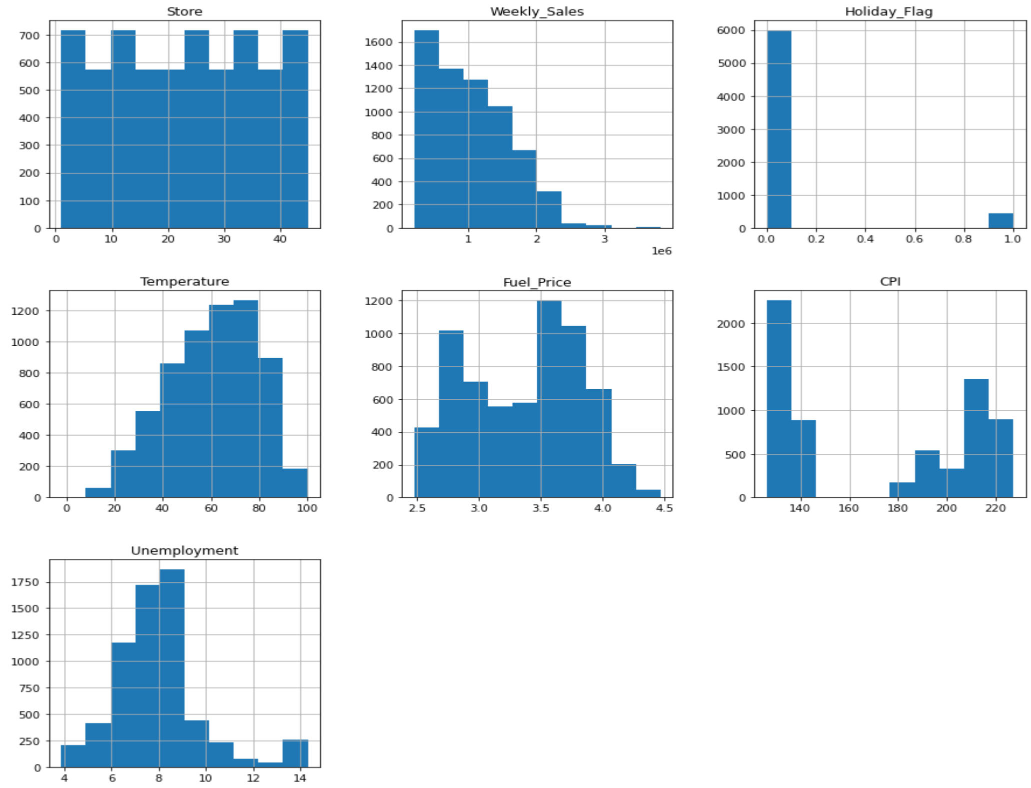


4. Weekly Sales and Unemployment

Another condition we can extrapolate from our dataset is the Unemployment rate.

In this graph we can see how Unemployment affects the weekly sales. We can conclude by saying that if the unemployment rate goes over 8 it has a negative impact on the weekly sales.

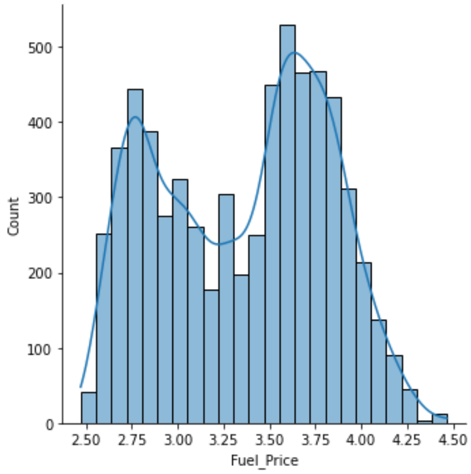




5. Weekly sales and Fuel Price

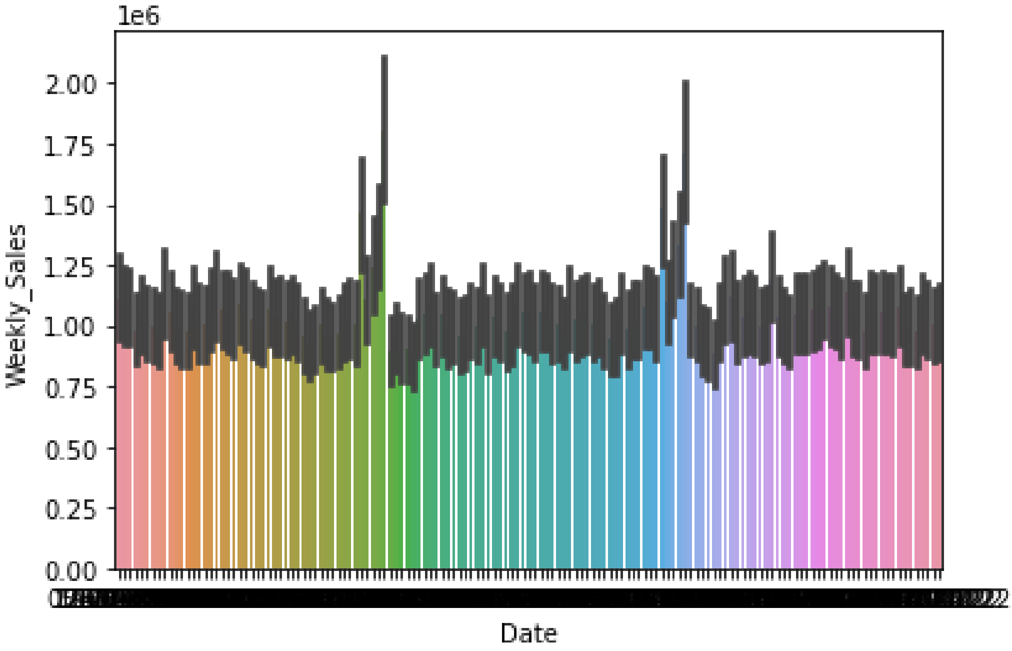
We thought it would be interesting to also see how the Fuel price impacts the sales.

Our reading of the data is the following, once the fuel price increases so do the weakly sales, but at one point they start decreasing again (after 3.75) the reasons could be multiple, but we think that since Walmart is considered a convenience store and the Fuel prices diverse base on areas, richer ares could prefer shopping in different stores. Another reason could be that period with a higher fuel price might encourage families to save more and this has an impact over weekly sales.



6. Weekly sales and Date

With this bar plot we can see the walmart sales through the dates. As we can see in the graph there are two peaks but in general the sales remain stable during all the dates, and there are no falls in the sales.



Data Manipulation

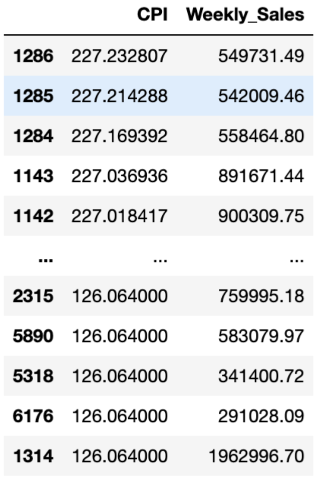
We thought it would be interesting to know how the CPI impacts the weekly sales. For this we made a sub set in order to better understand our data and have a more clear visualization of it.

```
cpi_data = walmart[['CPI', 'Weekly_Sales']]
cpi_data
```

Before analyzing the data we thought it is necessary to explain how the CPI values need to be read. A higher CPI means that consumer prices are higher, and when it falls it means consumer prices are generally falling. In short, a higher CPI indicates higher inflation, while a falling CPI indicates lower inflation, or in some cases deflation.

From the Analysis we can see how a lower CPI will drive our sales up.

cpi_data.sort_values(by = 'CPI', ascending = False)



	CPI	Weekly_Sales
1286	227.232807	549731.49
1285	227.214288	542009.46
1284	227.169392	558464.80
1143	227.036936	891671.44
1142	227.018417	900309.75
...	...	...
2315	126.064000	759995.18
5890	126.064000	583079.97
5318	126.064000	341400.72
6176	126.064000	291028.09
1314	126.064000	1962996.70

After we thought It would be interesting to see how the data are divided weekly,monthly, and yearly. By manipulating our data-set we were able to subtract that information and create a Boxplot for visualization for each of the three categories.

We chose a boxplot in that way we can better confront the data with the mean value. If the mean is higher it means that in the selected day, month or year, the sales were higher as well.

Boxplot Weekly

First, we manipulated our data-set to see how data are distributed during the week.

Legend:

1 - Monday

2 - Tuesday

3 - Wednesday

4 - Thursday

5 - Friday

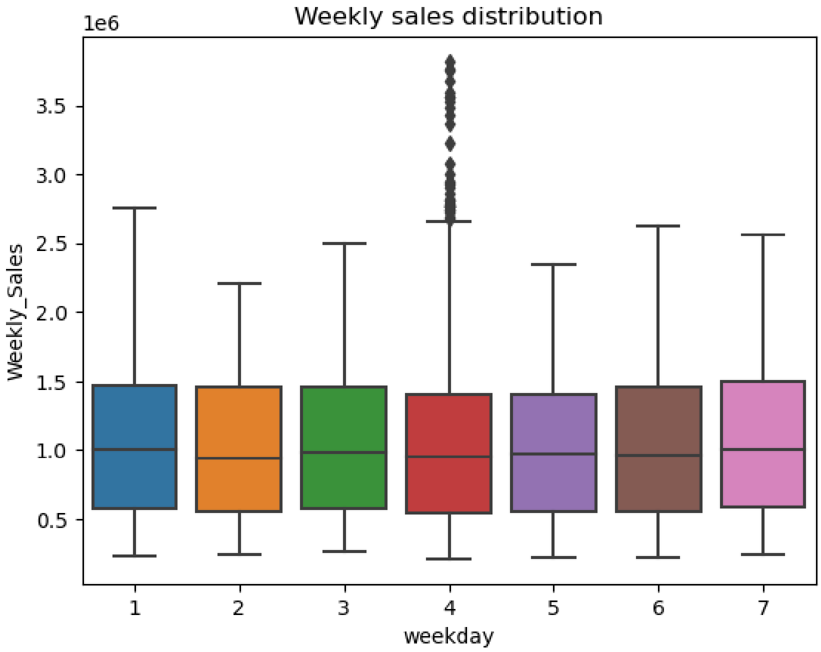
6 - Saturday

7 - Sunday

Weekday

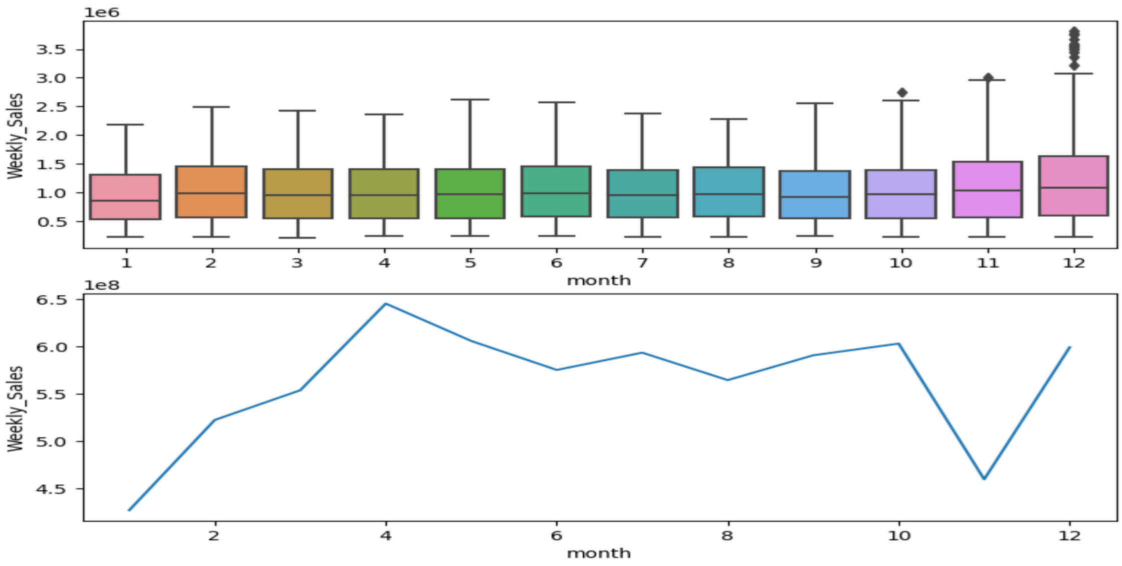
Afterwards, we decided to manipulate our data-set one more time to see how data is distributed during the weekday.

By looking at the graph we can say that: Monday(1), Wednesday(3), Thursday(4), Friday(5) are symmetric but Tuesday(2), Saturday(6) and Sunday(7) are positively skewed to the top.



Monthly

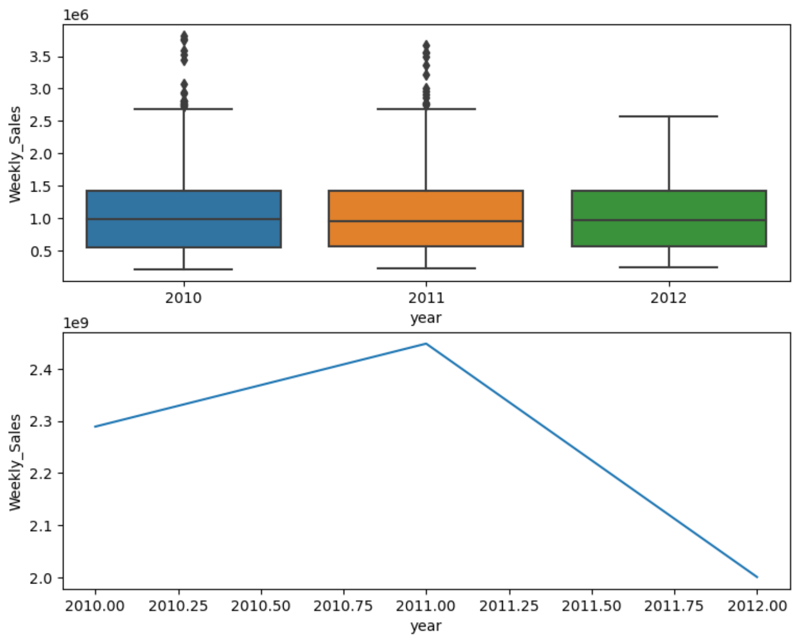
Monthly distribution of the dataset gave us a clear view of some months which have vital celebrations like Valentine's Day at February 14, Thanksgiving in November through to the Christmas celebration period. They all have a positive skewness in those months.



Yearly

In conclusion, we decided to manipulate our data-set in order to see how data are distributed during the years. We have the following years in consideration: 2010-2011-2012.

This information will also help us to better understand our prediction model.



Methodology/Model Building/Analysis

We started our project by a general explanation of what our purposes and objectives are, and explaining how our data set looks like.

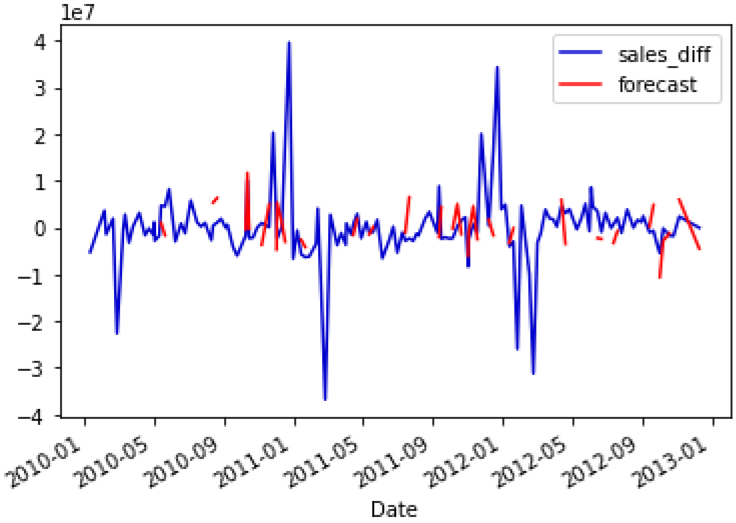
First we cleaned the dataset of any outliers and Na, and we had a general vision of what Walmart sales were and how they were distributed.

Afterwards we created different data visualizations that allowed us to understand how the different variables were connected to each other and how they were correlated in between. This allowed us to have a better vision and understanding of our dataset.

In this section of our project we will build a prediction model, this model should be able to tell us future predictions about Walmart sales.

We are using time series with the seasonal ARIMA model, also the ARIMA model itself since we don't have seasonality in our data set.

Based on the graph we can analyze the first month of 2013 as a prediction, they are not far from what was happening the previous years. In January, sales were pretty stable and after it is expected a slight decrease till April. This is exactly what happened during the two previous years so it is correct to have similar expectations. Looking more deeply into it we came up to a suitable conclusion; sales are pretty stable and higher when holidays are expected: December and January. People are usually more likable to spend money over the holidays. Once holidays end, so do the different deals. This might cause the decrease we assist after january. It is not a coincidence that sales go up again around April, it is a result of Easter break. Walmart puts in place different deals and customers are once again attracted to shopping. Holidays also bring a positive spending effect in customers mind, since they feel more secure about spending money.



Conclusion/Recommendations

After analyzing the filtered and sorted data we can conclude that there are some significant correlations between different factors. Using the Spearman's rank we could see CPI and unemployment and weekly sales with stores have a high correlation. From the visualizations we created, we focused on how sales are affected by special weeks such as holidays, but also depending on the temperature, and other factors such as CPI, unemployment and fuel price. From the pie chart we created, we saw that 7.5% of the overall sales were made during special holiday weeks. This means that Walmart generates a lot of sales during the holiday season. From one of the histograms we made, we can see that sales rise when the temperatures are higher. Another factor we took into account is the unemployment rate. From the histogram we could see that the sales were negatively affected when the rate of unemployment was over 8%. Fuel price also plays an important role, as we saw that when the price started to rise, sales decreased. However, it was interesting to see that when the fuel price started to rise even more sales eventually increased with it. Nevertheless, when the fuel price was over $3.75/gallon the sales fell again. From some of the data manipulation we created a subset of columns to analyze how CPI affected sales. We could see that a lower CPI increased sales. We also created a boxplot weekly. We manipulated the data set and created weeks and after the days of the week, Monday through Sunday. Thanks to the boxplot we can see that there are four days (Monday, Wednesday, Thursday, and Friday) that are symmetric and the rest of the days are skewed to the top.

By looking at the monthly graph we can see that there are certain months with important celebrations that stand out in the graph, such as February, with Valentine's Day, and the Christmas months also, especially the December peak. In conclusion, we have modified the data for the years 2010, 2011, 2012 and thanks to that we can better understand the prediction model. As for last, we decided to create a prediction model for what the sales will be for the upcoming year, 2013. We looked only into the first month of 2013 and if we compared the date of January sales they are not really different but a slight decrease is expected for the upcoming trimester. After our analysis we saw how sales are directly influenced by the economic environment, and holiday weeks.

Works Cited

Kaggle - Walmart

https://en.wikipedia.org/wiki/Walmart

https://www.tutor2u.net/business/reference/sales-forecasting

https://www.demandjump.com/blog/what-are-the-factors-affecting-sales-forecasting