**WALMART WEEKLY SALES FORECASTING**

**INTRODUCTION**

The marketing department for Walmart is interested in determining the optimal allocation of marketing resources for future growth and to manage cash flow. Sales forecasting will help Walmart set benchmarks for future trends and allow Walmart to course correct and optimize early wins.

There are various things that influence the weekly sales like **product promotions, sale markdowns, weather, competitor actions, festive demands,** etc. For example, when its warmer people spend more time outdoors. It is likely that outdoor wear sales would increase during the summers then the winters.

**DATA OVERVIEW**

Data was collected from year **2010–2012** for**45 Walmart stores**. We are tasked with predicting sales for the next year. We are provided with datasets: Stores.csv,train.csv,test.csv,fetaures.csv

**Dataset Size**

**Trainset :**421570 rows (12.2MB)

**Testset :** 115064 rows (2.47MB)

**Stores :** 4KB

**Features:** 580 KB

**Column Descriptions**

**Stores :**

**Type:** Type of the store namely “A”,” B”,”C”

**Size**: Size of the store. Size refers to the number of products inside the store. Range 34000 -210,000 8/4/2021 4/64

**Sales (Trainset/Testset):**

***Store*** — The store number. Range 1- 45.

***Dept*** — One of 1–99 that shows the department

***Date*** — the week

***Weekly Sales*** — sales for the given department in the given store

***IsHoliday*** — whether the week is a special holiday week.

**Features:**

**Temperature** — Average temperature in the region during that week.

**Fuel\_Price** — cost of fuel in the region during that week

**MarkDown1–5 —** anonymized data related to promotional markdowns that Walmart is running. Markdown data is only available after Nov 2011 and is not available for all stores all the time.

Any missing value is marked with an NA. represents what quantity was available during that week. Type is from 1–5.

**CPI**— the consumer price index during that week.

**Unemployment** — the unemployment rate during that week

# EXPLORATORY DATA ANALYSIS

* CPI, Unemployment and Markdown had null values and we replaced them with mean. We see that few stores have missing departments and dates and their reported weekly sales. We marked their weekly sales as ‘0’ and 'ISholiday' column as ‘0’ for No Holiday and ‘1’ for Holiday.
* We are going to treat missing values by two ways:
* If there are more than 75% of null values, we will drop the column as we can’t even retain 75% information from them.
* If there are less than 75% of null values, we will perform imputation such as replacing with mean value per store.

Text, table

Description automatically generated

We didn’t have any feature that had missing values greater than 75% therefore we didn’t drop any column.

We checked for duplicates in our data and didn’t find any.

Chart, bar chart

Description automatically generated

From the bar chart above, Thanksgiving Holiday has the highest Average Sales compared to other types of Holiday.

Chart, bar chart

Description automatically generated

From this bar chart, Store Type A has the most average Weekly Sales for every Holiday Type. This could be due to the size and the products available at Store Type A.

Chart, pie chart

Description automatically generated

From this pie chart, store Type A has higher sales > Type B > Type C

Chart, line chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Observations**: We can observe that during thanksgiving week(Nov 26) and christmas week(December 24) soaring peaks of sales.Holidays like superbowl occurs in first sunday of feb month,labour day on first monday of September.

**Chart, bar chart

Description automatically generated**

Few stores show higher mean sales and higher interquartile range showing that some stores are bigger than others

Some stores have their 25th and 50th percentile values very close and 75th percentile is far away showing that there is not much sales values difference between point below 25% and point below 50%

**Theory :** We can understand that the stores with the higher weekly sales(1,2,4,10...) can be categorized as type 'A' stores with medium sales as type 'B'(11,2,19,20...) and the rest as type 'C'(3,5,29,30...)

Chart

Description automatically generated

From this box plot we can see department 2,38,40,92 and 95 has greater mean sales comparated to other departments which shows that these department have more useful products for the customer.People are more likely to buy products from these departments. Department 72,7,5 we can observe high sales during few weeks which is farther away from their mean sales.There is a posibility that these departments supply customers with products which are needed during Holiday Weeks.

Chart, line chart

Description automatically generated

During the holidays, there is peak in sales for the last two months for the stores A and B(the bigger stores)

Chart, box and whisker chart

Description automatically generated

The average weekly sales is the same accross all four quarters. The last quarter has a peak in sale in some week as the max sale for the last quarter is the highest.

**Observations**: As we can see outlier points on Holidays are greater than non-Holidays.The mean sales/inter-quartile range values for both Holiday and nonholiday weeks are overlapping and this **doesn’t make ‘IsHoliday’ a good predictor**. Which indicates are only a few Holiday weeks which clearly increases sales values than non-Holiday Weeks.

Chart, bar chart, line chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

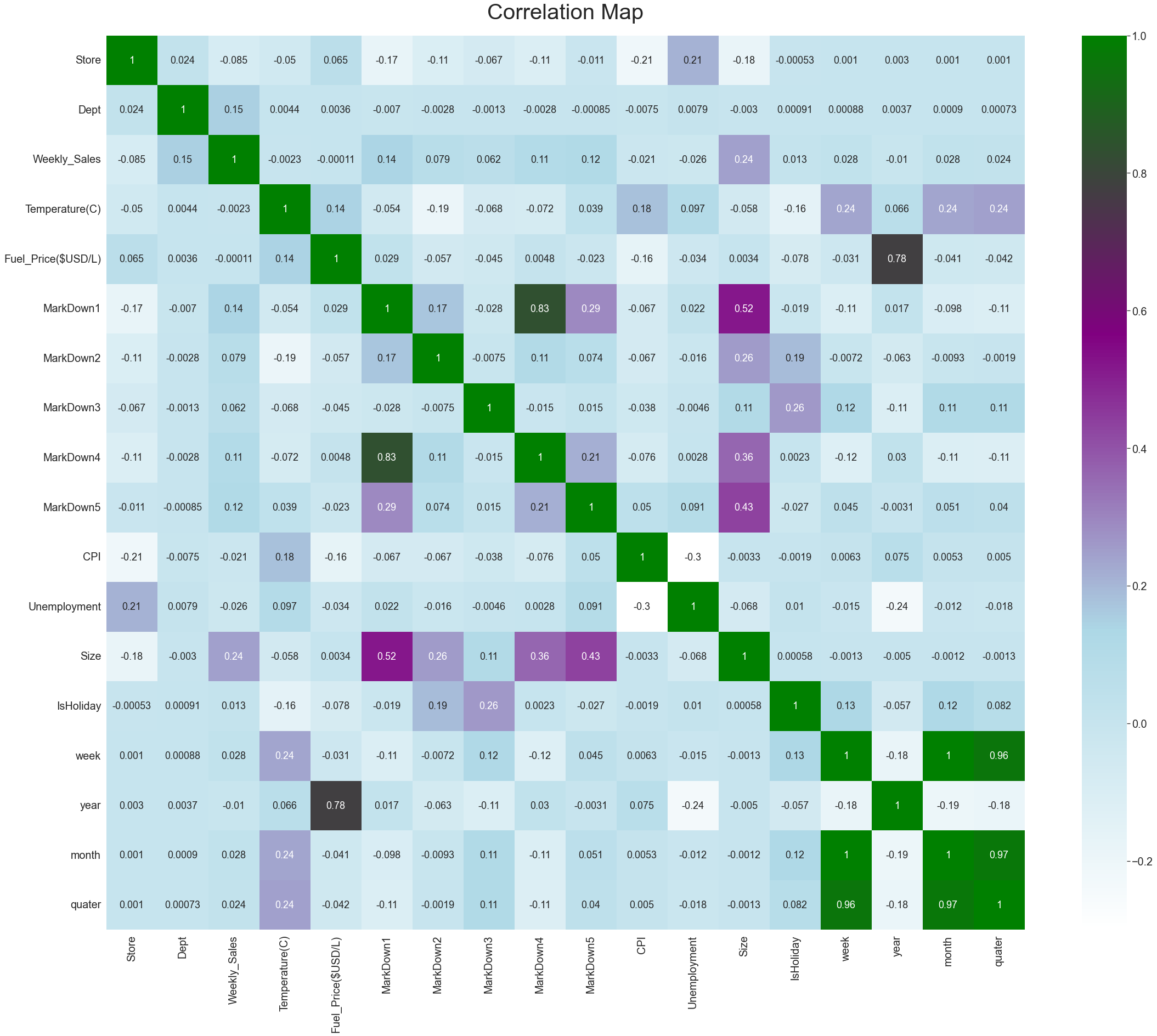
There isn't any strong correlation between weekly sales and CPI, Unemployment rate, Temperature(C) and Fuel Price

Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated



Fuel price and year are highly correlated,hence we can decide to drop Fuel\_price. MarkDown1 and MarkDown4 are corelated ,lets drop markdown 4.**Size has the strongest corelation(0.2) with Weekly\_Sale**s.Fuel\_price and Temperature has very weak coorelation with sales.Week is strongly co-related with month,dayofyear and quater,**let’s drop column dayofyear and month as it’s strongly corelated to week**.

**DATA MODELLING AND UNIVARIATE TIME SERIES ANALYSIS**

Using time-series we can get the insight of how time can add weightage in determining the value of an asset/commodity or variable of our concern. In our ususal Machine Learning analysis we check how the independent variables affecting our dependent variable,order doesn’t matter. In our time series analysis we check how the past observations of our target variables varies with respect to time .Here the data is ordered with respect to time which creates a dependency between past and future observations.

Graphical user interface

Description automatically generated

**Observations**: We can see clearly that data contains both upwards and downward trends.**Additive Seasonality** — We can see that there is 90% spike in sales during year-end every year additively. Data is also quite **noisy**, which depicts uncertainty in sales in some periods . Here is when we use the smoothening technique. We are going to average out some past values which will eventually reduce noise in data.

**Tests for Stationarity**

A time series is *stationary* if the mean and variance are fixed between any two equidistant points. That is, no matter where you take your observations, the results should be the same. A times series that shows seasonality is *not* stationary.

A test for stationarity usually involves a unit root hypothesis test, where the null hypothesis H0 is that the series is *nonstationary*, and contains a unit root. The alternate hypothesis H1 supports stationarity. The augmented Dickey-Fuller and Kwiatkowski-Phillips-Schmidt-Shin tests are stationarity tests.

**Augmented Dickey-Fuller Test**

To determine whether a series is stationary we can use the augmented Dickey-Fuller Test. In this test the null hypothesis states that ϕ=1 (this is also called a unit test). The test returns several statistics we'll see in a moment. Our focus is on the p-value. A small p-value (p<0.05) indicates strong evidence against the null hypothesis.

For our data, the Augmented Dickey Fuller Test proves that the data is stationary since the p-value is less than 0.05.

## **Forecasting with the Holt-Winters Method**[**¶**](http://localhost:8888/notebooks/Downloads/Walmart/Data%20Modeling.ipynb#Forecasting-with-the-Holt-Winters-Method)

We decided to forecast the sales one year ahead. The test set was divided in such a way that the last 52 weeks were set for the test set.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

We used RMSE as Evaluation Metrics and found the RMSE as 238.0810335790535.

**SARIMA(p,d,q)(P,D,Q)m**

**Seasonal Autoregressive Integrated Moving Averages**

Where ARIMA accepts the parameters (𝑝,𝑑,𝑞)(p,d,q), SARIMA accepts an *additional* set of parameters (𝑃,𝐷,𝑄)𝑚(P,D,Q)m that specifically describe the seasonal components of the model. Here 𝑃P, 𝐷D and 𝑄Q represent the seasonal regression, differencing and moving average coefficients, and 𝑚m represents the number of data points (rows) in each seasonal cycle.

Chart, line chart

Description automatically generated

SARIMA(2,0,2)(1,0,1,52) RMSE Error: 773.9530957. The ARIMA component of this model was 2,0,2 and the seasonality component was (1,0,1,52). We chose m as 52 because of 52 weeks in a year and we wanted to forecast sales one year ahead.

Chart

Description automatically generated

**CONCLUSION:**

Holt Winters Method is a better model for forecasting the sales since it has lower RMSE. This data was stationary and didn’t have any seasonality therefore Hodrick Prescott Filter seems to be a better model for forecasting the sales.

We can use LSTM in the future to capture the sales as this is a MultiVariate Time Series Analysis.