2020

Analysis of Walmart Sales & Sales Forecast

MILESTONE REPORT

NGOC PHAN

Project Background

According to Wikipedia, "Walmart is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores". The company has 45 stores across the United States. Every year the company runs several promotional markdown events to increase sales. These markdowns precede prominent holidays such as Super Bowl, Labor Day, Thanksgiving, and Christmas.

Problem Statement

This project studies Walmart's historical sales data for 45 stores in the United States to assist management team in identifying factors that affect sales and forecasting future sales by:

- Performing exploratory data analysis and time series analysis of Walmart's sales data
- Developing machine learning algorithms to forecast sales

Dataset

Walmart's sales datasets are collected on Kaggle website at https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data. The datasets contain historical sales data for 45 Walmart stores in the United States along with store information and regional activity from 2/5/2010 to 11/1/2012. There are 3 csv files: stores, train, and features. The variables are described below:

- **stores.csv**

  This file contains information about 45 Walmart stores and includes the following fields:
  - ***Store*** – the store number
  - ***Type*** – the type of store (A, B, or C)
  - ***Size*** – the size of store in square feet
- **train.csv**

  This file contains Walmart historical sales data from February 5th, 2010 to November 1st, 2012 and includes the following fields:
  - ***Store*** – the store number

- ***Dept*** – the department number
- ***Date*** – last day of the week
- ***Weekly_Sales***
  - Weekly sales for the given department in the given store
  - Negative if returns exceed sales
  - Positive if sales exceed returns
- ***IsHoliday*** – True if special holiday falls within the week; otherwise, False

- **features.csv**

  This file contains data related to the store, department, and regional activity for the given dates and includes the following fields:

  - ***Store*** – the store number
  - ***Date*** – last day of the week
  - ***Temperature*** – average temperature in the region in Fahrenheit
  - ***Fuel_Price*** – weekly average fuel price (USD)
  - ***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.
  - ***CPI*** – consumer price index
  - ***Unemployment*** – weekly average unemployment rate
  - ***IsHoliday*** – True if holiday falls within the week. False if holiday does not fall within the week.

Data Preparation

Missing Values

There are seven columns that contain missing values including MarkDown 1-5, CPI, and Unemployment. Since most missing values exist because there was no information available at one specific time, fields containing missing values are left as 'NA'. The table below lists columns that have missing values along with statistics:

table	col	null_count	null_pct	min	max	mean	median
features	MarkDown1	4158	51	-2781.0	103185.0	7032.0	4744.0
features	MarkDown2	5269	64	-266.0	104520.0	3384.0	365.0
features	MarkDown3	4577	56	-179.0	149483.0	1760.0	36.0
features	MarkDown4	4726	58	0.0	67475.0	3293.0	1176.0
features	MarkDown5	4140	51	-185.0	771448.0	4132.0	2727.0
features	CPI	585	7	126.0	229.0	172.0	183.0
features	Unemployment	585	7	4.0	14.0	8.0	8.0

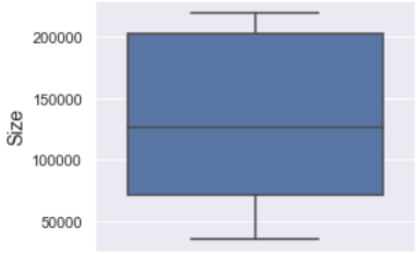
For 2010 sales data, we are missing the observations for the first four weeks of the year. Therefore, new observations have been added to the dataset for the first four weeks of 2010 using the 2011-2012 average sales of the corresponding period.

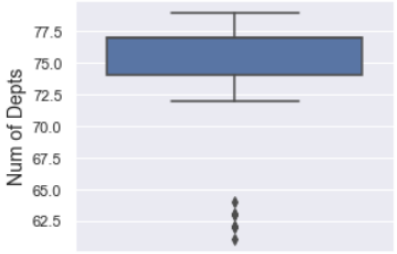
New Columns

- Num of Depts = counts of number of departments for each store
- Weekly Sales (Thousands) = Weekly Sales / 1000
- Total Sales (Millions) = total Weekly Sales in Thousands / 1000
- Total MarkDown = sum of MarkDown1-5
- Total MarkDown (Thousands) = Total MarkDown / 1000
- Year = year extracted from Date
- Quarter = quarter extracted from Date
- Month = month extracted from Date
- Week = week of year extracted from Date

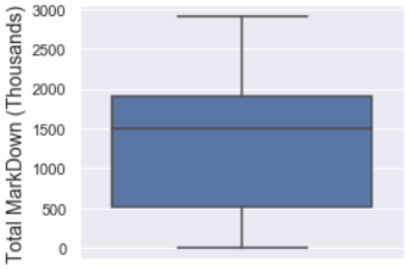
Outliers

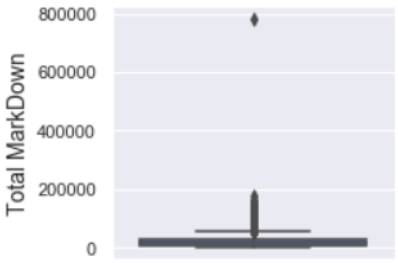
According to the box plots below, there are 11 columns that have outliers: Num of Depts, Weekly_Sales, Temperature, MarkDown1-5, and Unemployment.

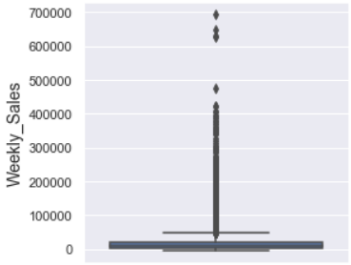


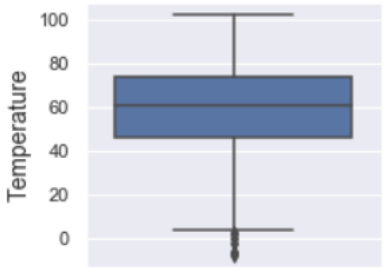


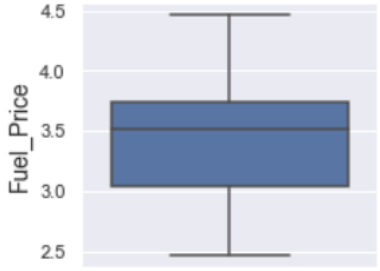

Total Sales (Millions)
300
250
200
150
100
50

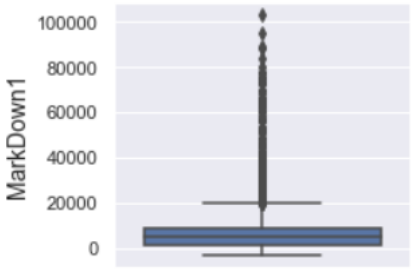

Total MarkDown (Thousands)
3000
2500
2000
1500
1000
500
0

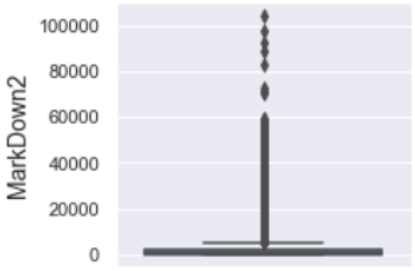

Total MarkDown
800000
600000
400000
200000
0

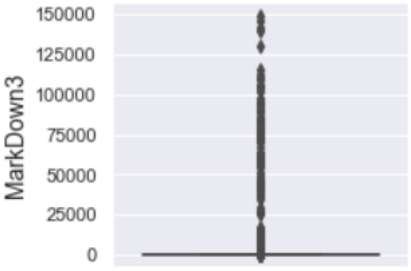

Weekly_Sales
700000
600000
500000
400000
300000
200000
100000
0

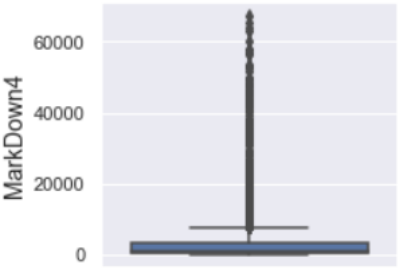

Temperature
100
80
60
40
20
0

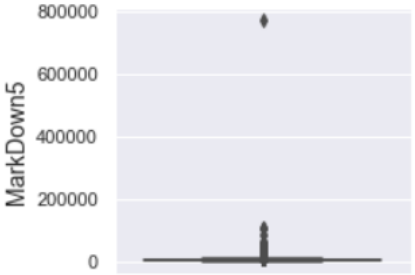

Fuel_Price
4.5
4.0
3.5
3.0
2.5

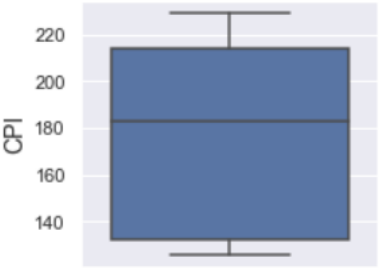

MarkDown1
100000
80000
60000
40000
20000
0

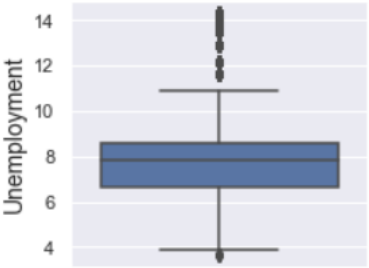

MarkDown2
100000
80000
60000
40000
20000
0


MarkDown3
150000
125000
100000
75000
50000
25000
0


MarkDown4
60000
40000
20000
0

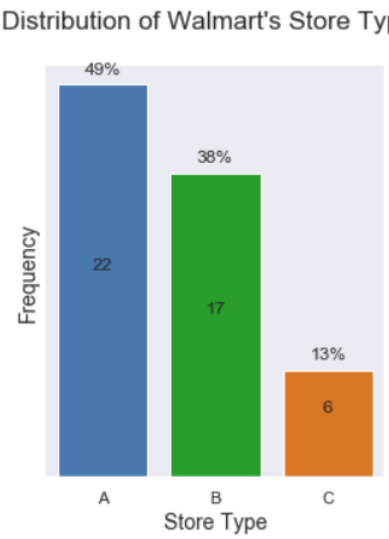

MarkDown5
800000
600000
400000
200000
0


CPI
220
200
180
160
140


Unemployment
14
12
10
8
6
4

Exploratory Data Analysis

Walmart stores are classified into 3 types: A (22 stores), B (17 stores), and C (6 stores).



The total sales for each Walmart store ranges between $37 million and $301 million. About half of the 45 Walmart stores have total sales greater than or equal to $138 million.

	Total Sales (Million)
count	45.000000
mean	149.715977
std	78.167556
min	37.160222
25%	79.565752
50%	138.249763
75%	199.613906
max	301.397792

The weekly sales for each Walmart's store ranges from $209 thousand to $3.8 million. About half of the 45 Walmart stores have weekly sales greater than or equal to $960 thousand.

	Weekly Sales (Thousand)
count	6435.000000
mean	1046.964878
std	564.366622
min	209.986250
25%	553.350105
50%	960.746040
75%	1420.158660
max	3818.686450

The weekly sales for Walmart's department store ranges from -$4,988 to $693,099. The negative values in weekly sales indicates returns exceed sales in the store's department. About half of the Walmart's department store have sales greater than or equal to $7,612.

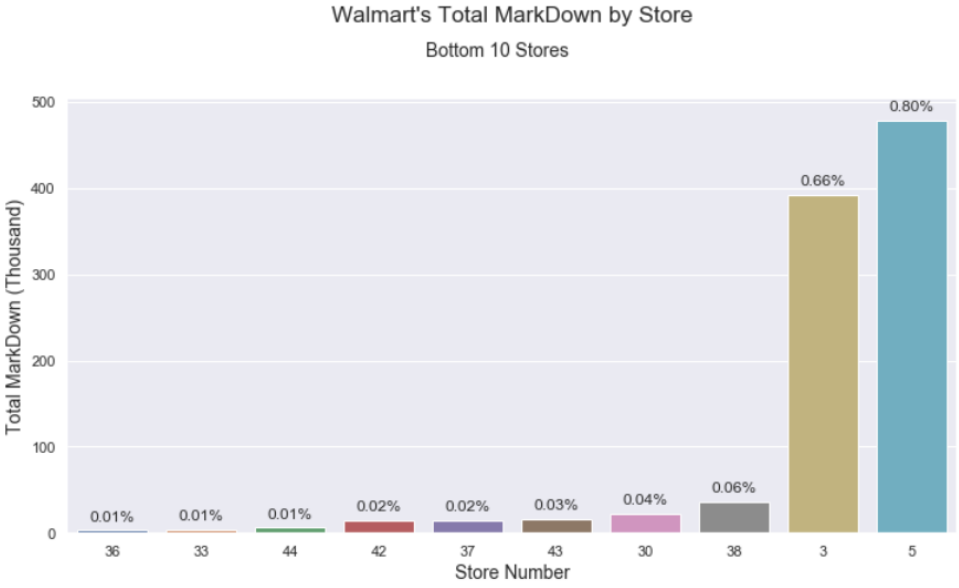
	Weekly_Sales
count	421570.000000
mean	15981.258123
std	22711.183519
min	-4988.940000
25%	2079.650000
50%	7612.030000
75%	20205.852500
max	693099.360000

Stores that are in in the top 10 sales contribute more than 3% (or more than $200 million) toward Walmart's total sales.

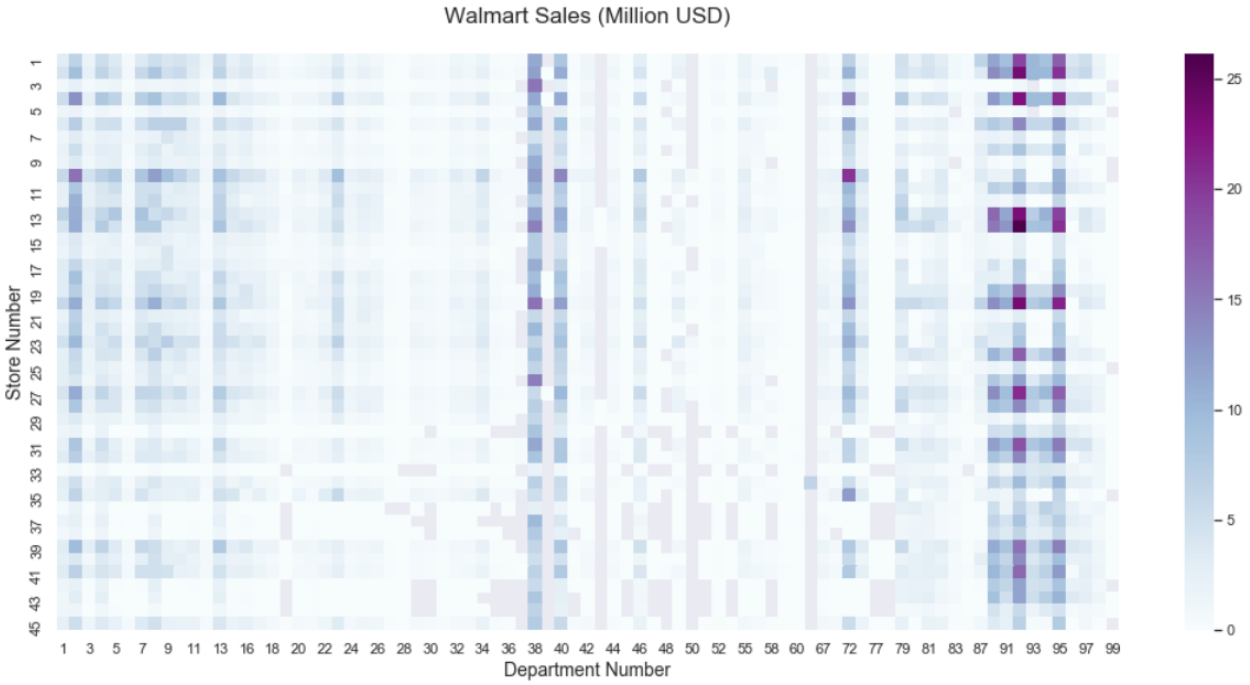
Stores that are in the bottom 10 sales contribute 1.15% or less (or $77 million or less) toward Walmart's total sales.

Seven stores in the top 10 sales are also in the top 10 markdowns. Those stores are 20, 4, 14, 13, 2, 27, and 39.

Seven stores in the bottom 10 sales are also in the bottom 10 markdowns. Those stores are 44, 5, 36, 38, 3, 30, and 37.



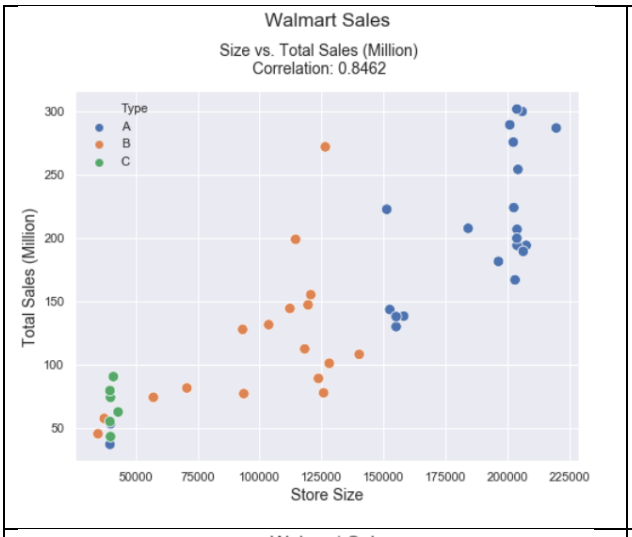
Departments 38, 40, and 88 to 95 in most stores seem to have higher sales than other departments.

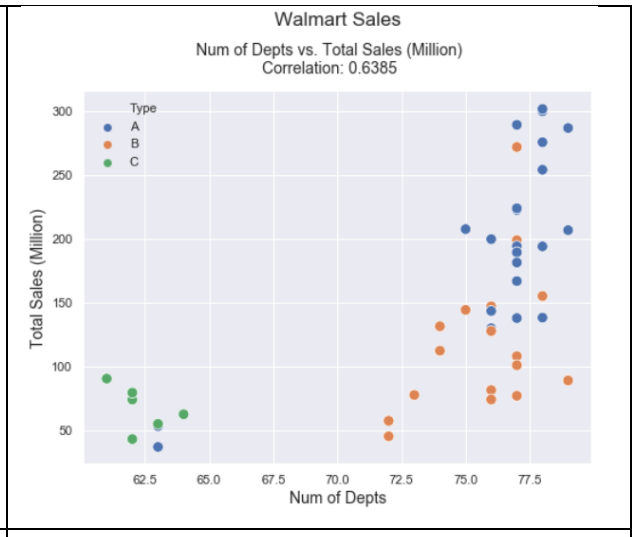


There is a positive linear relationship with strong correlation between the following variables:

- Store's size and total sales
- Number of departments in store and total sales
- Total markdowns and total sales

Most type A stores seems to be in a group that have highest total sales. Most type C stores appears to be in a group that have lowest total sales. Most type B stores' total sales are higher than type C stores' total sales and lower than type A stores' total sales.





The heatmap bellow shows that there isn't a strong correlation between the following variables:

- Weekly sales and temperature
- Weekly sales and fuel price
- Weekly sales and CPI
- Weekly sales and unemployment
- Weekly sales and total markdown

Time Series Analysis

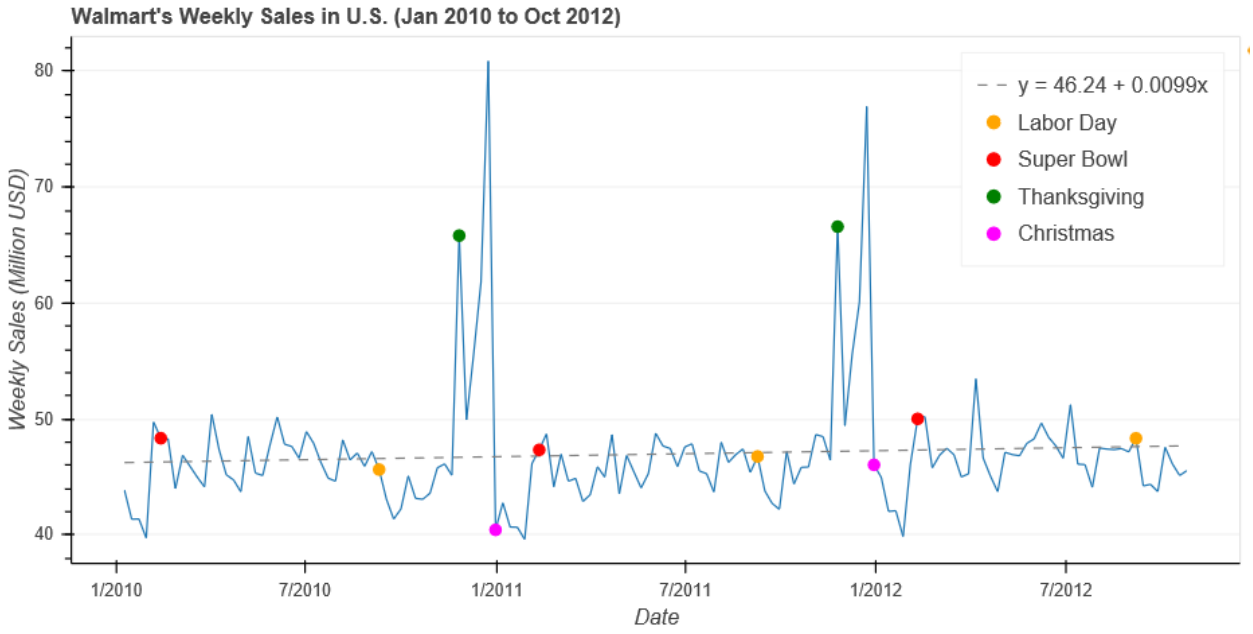
Original Time Series and General Trend

Holidays do not seem to have much impact on Walmart's weekly sales except for Thanksgiving. Sales appear to be high during and after the weeks of Thanksgiving. The general trend line (dashed gray line) seems to be flat and has the following equation:

$$y = 46.24 + 0.0099x$$

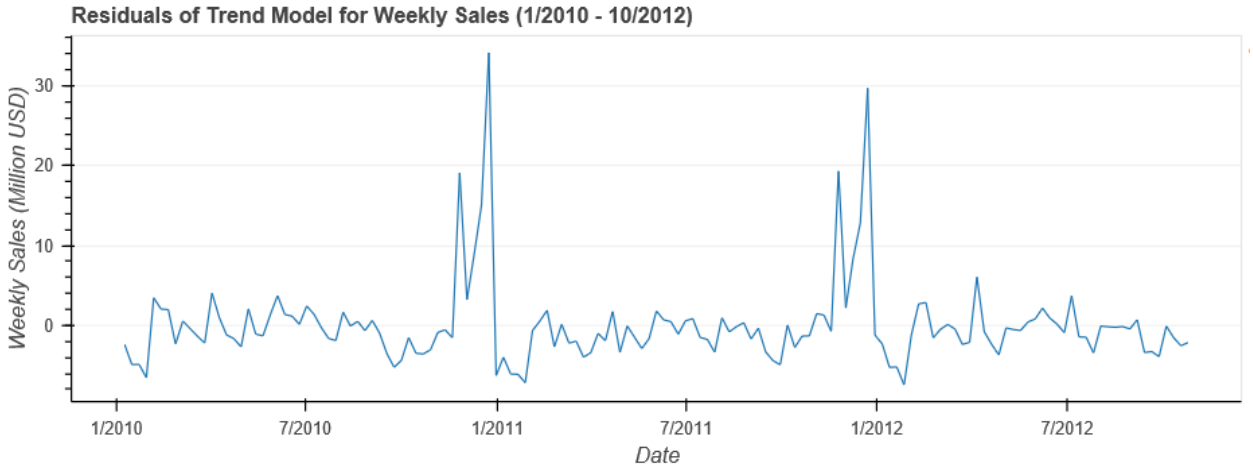
where

- ***y*** = Weekly Sales (Million)
- ***x*** = date index (e.g. first week has index of 1, second week has index of 2, etc.)

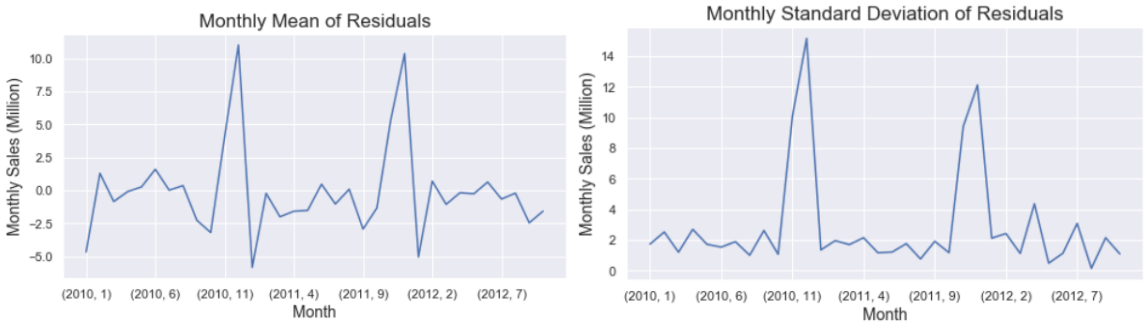


Seasonality

Residuals plot of general trend model indicates that sales data exhibit seasonality because there are spikes in sales during and after the weeks of Thanksgiving.



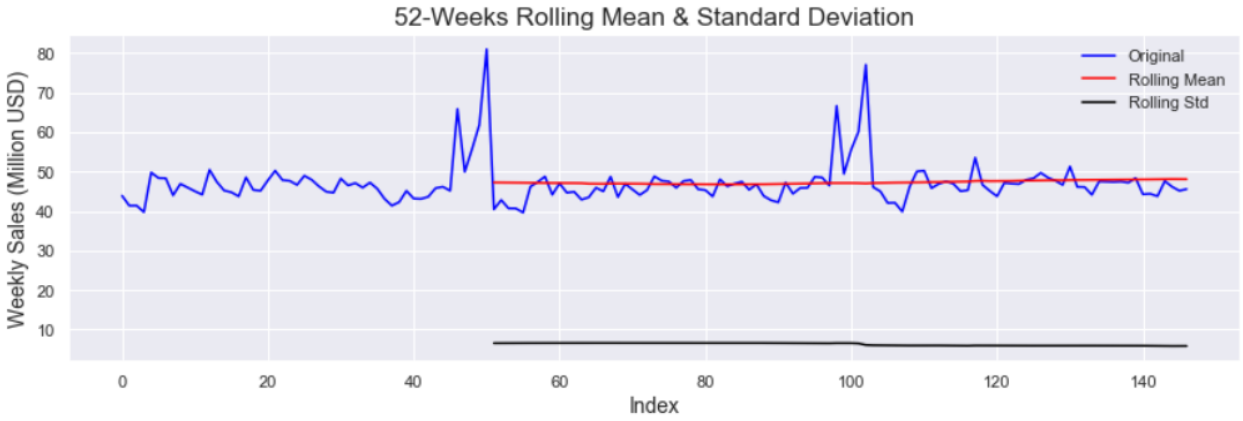
The Monthly Mean of Residuals plot and Monthly Standard Deviation of Residuals plot also indicate seasonality between November and December.



The box plot of Walmart's quarterly sales shows that sales are almost the same for quarter 1 to quarter 3. Sales are highest for fourth quarter of 2010 and 2011. Since 2012 sales data does not include sales for November and December, sales for fourth quarter of 2012 are almost the same as sales for other quarters.

Stationarity

The plot of 52-weeks rolling mean and rolling standard deviation of weekly sales shows that the data fluctuate around the mean.



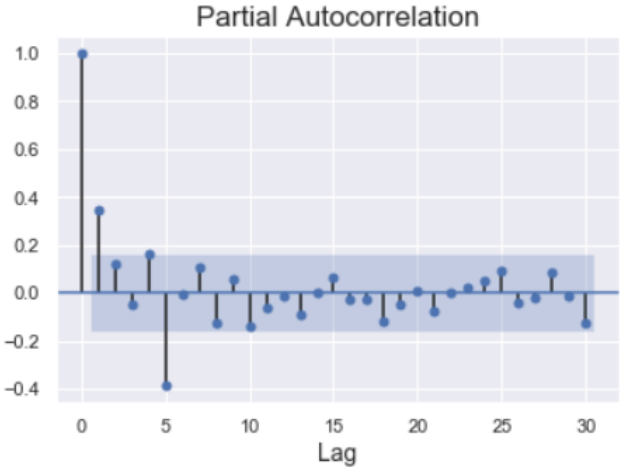
The result of Dickey-Fuller test also indicates time series is stationary because the p-value is less than 0.05.

```
Results of Dickey-Fuller Test:
Test Statistic                -5.977907e+00
p-value                        1.868362e-07
#Lags Used                     4.000000e+00
Number of Observations Used    1.420000e+02
Critical Value (1%)           -3.477262e+00
Critical Value (5%)           -2.882118e+00
Critical Value (10%)          -2.577743e+00
```

Model Development for Sales Forecast

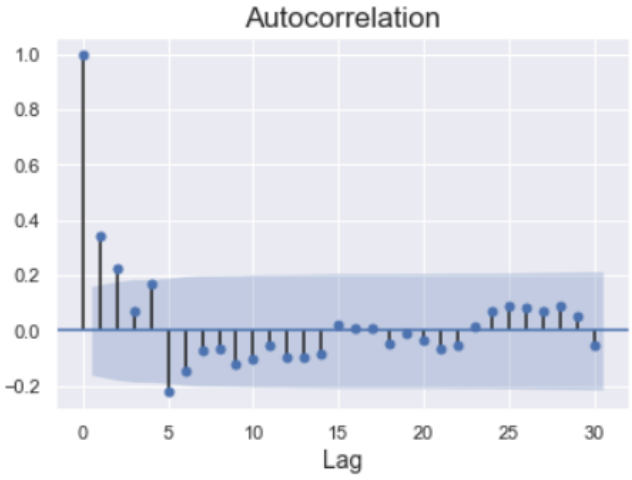
Autoregressive (AR) Term

The partial autocorrelation plot indicates that lag 1 and 5 are the best orders for AR term.



Moving Average (MA) Term

The autocorrelation plot indicates that lag 1, 2, and 5 are the best orders for MA term.



ARMA Model

Since the time series is stationary, the order for differencing is 0. The figure below shows the model's summary for ARMA(1, 2). The p-value of the coefficients are all significant (p-value less than 0.05) for the constant and all AR and MA levels.

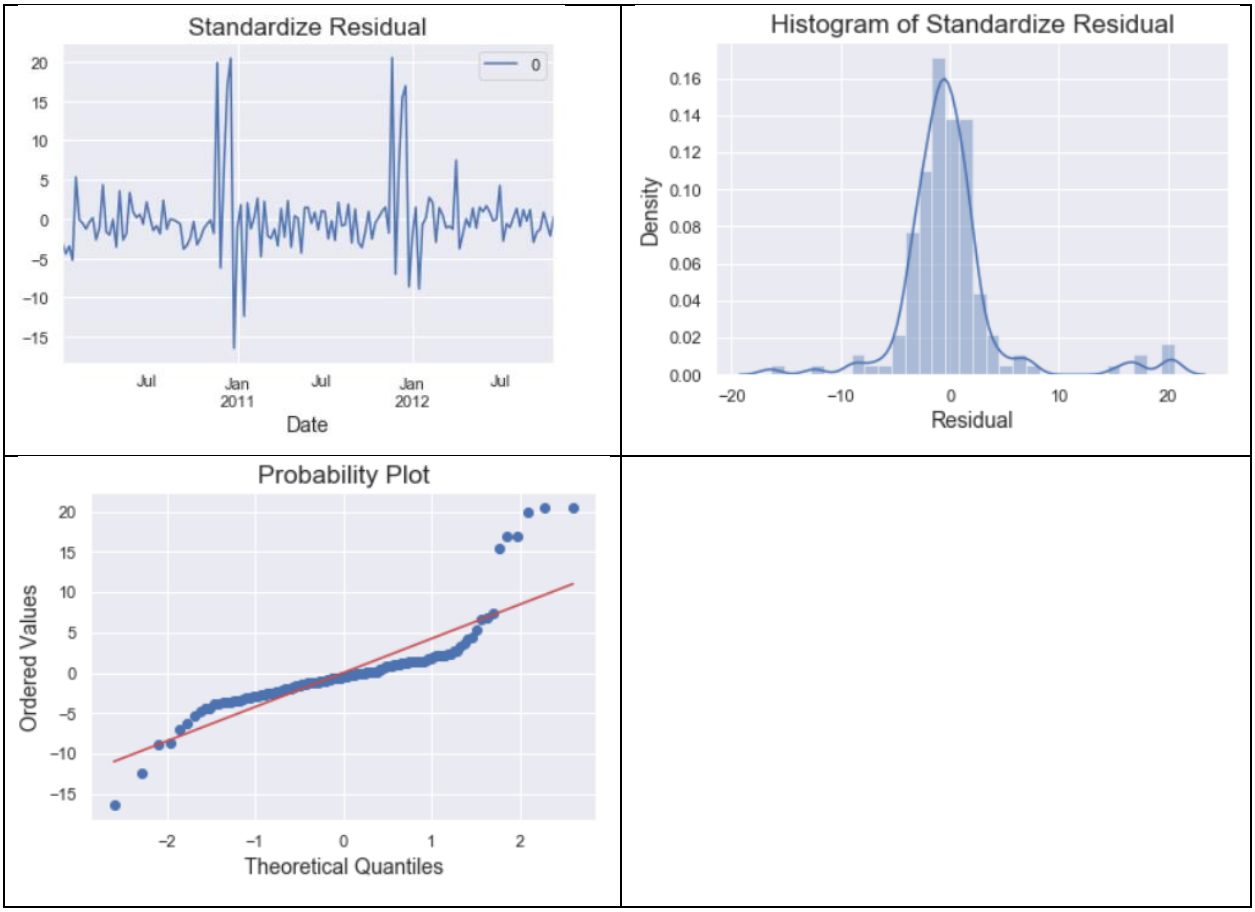
Dep. Variable:	Weekly Sales (Millions)	**No. Observations:**	147
Model:	ARMA(1, 2)	**Log Likelihood**	-440.155
Method:	css-mle	**S.D. of innovations**	4.819
Date:	Tue, 21 Apr 2020	**AIC**	890.310
Time:	20:14:29	**BIC**	905.262
Sample:	01-08-2010	**HQIC**	896.385
	- 10-26-2012		

	coef	std err	z	P>|z|	[0.025	0.975]
const	46.9451	0.642	73.106	0.000	45.687	48.204
ar.L1.Weekly Sales (Millions)	-0.7320	0.086	-8.509	0.000	-0.901	-0.563
ma.L1.Weekly Sales (Millions)	1.2129	0.084	14.509	0.000	1.049	1.377
ma.L2.Weekly Sales (Millions)	0.5935	0.087	6.837	0.000	0.423	0.764

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.3662	+0.0000j	1.3662	0.5000
MA.1	-1.0219	-0.8005j	1.2981	-0.3942
MA.2	-1.0219	+0.8005j	1.2981	0.3942

The mean of residuals are closed to zero. The residuals are also fluctuated around the mean. It seems that ARMA(1, 2) model is not a good fit because the model does not take seasonality into account. The probability plot shows that the points do not hug the red line closely. There are also some extreme values at the top right corner and bottom left corner of the probability plot.



The table below shows the metric for ARMA(1, 2) model. The mean squared error is around $23 million dollars. The Mean Absolute Percentage Error (MAPE) of 0.055867 indicates that the model is about 94.5% accurate in predicting the next 52 observations.

mape	me	mae	mpe	mse	rmse	corr	minmax
0.055867	-0.004523	2.789844	0.007588	23.296938	4.82669	0.463302	0.05262

The plot below shows the model's observed and forecast sales.

SARIMAX Model

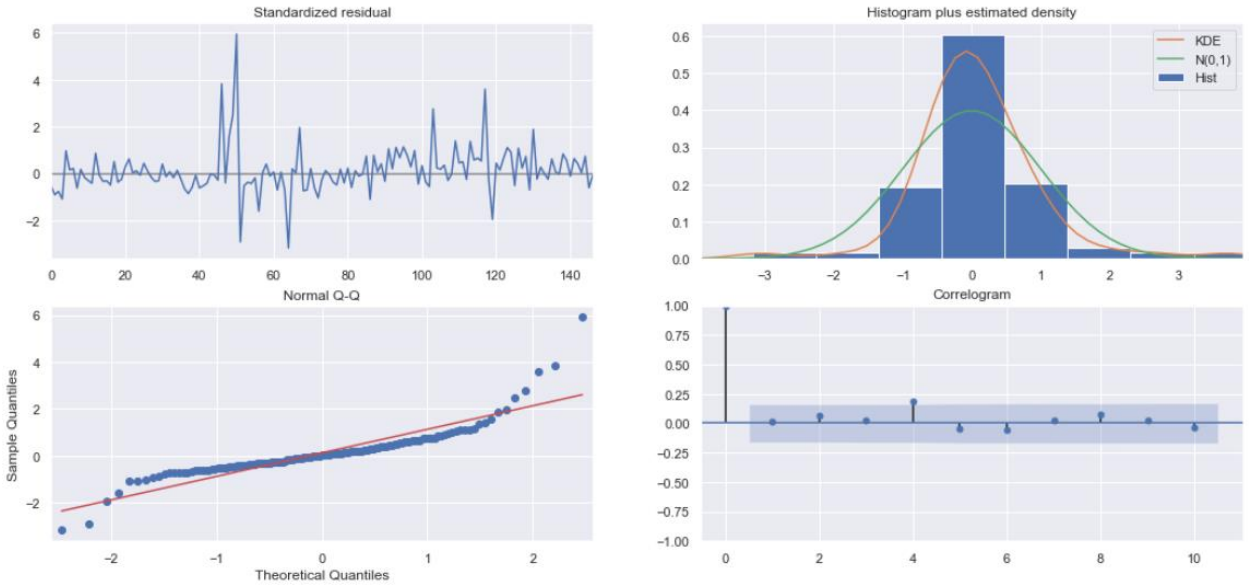
The figure below shows the model's summary for SARIMAX model. The p-value of the coefficients are all significant.

Dep. Variable:	y	**No. Observations:**	147
Model:	SARIMAX(1, 0, 0)x(1, 0, 0, 52)	**Log Likelihood**	-359.917
Date:	Tue, 21 Apr 2020	**AIC**	727.834
Time:	20:48:05	**BIC**	739.796
Sample:	0	**HQIC**	732.694
	- 147		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	2.6581	0.431	6.161	0.000	1.813	3.504
ar.L1	0.2734	0.052	5.254	0.000	0.171	0.375
ar.S.L52	0.9217	0.009	98.294	0.000	0.903	0.940
sigma2	3.9129	0.379	10.330	0.000	3.170	4.655

Ljung-Box (Q):	26.05	**Jarque-Bera (JB):**	610.89
Prob(Q):	0.96	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.61	**Skew:**	1.71
Prob(H) (two-sided):	0.10	**Kurtosis:**	12.39

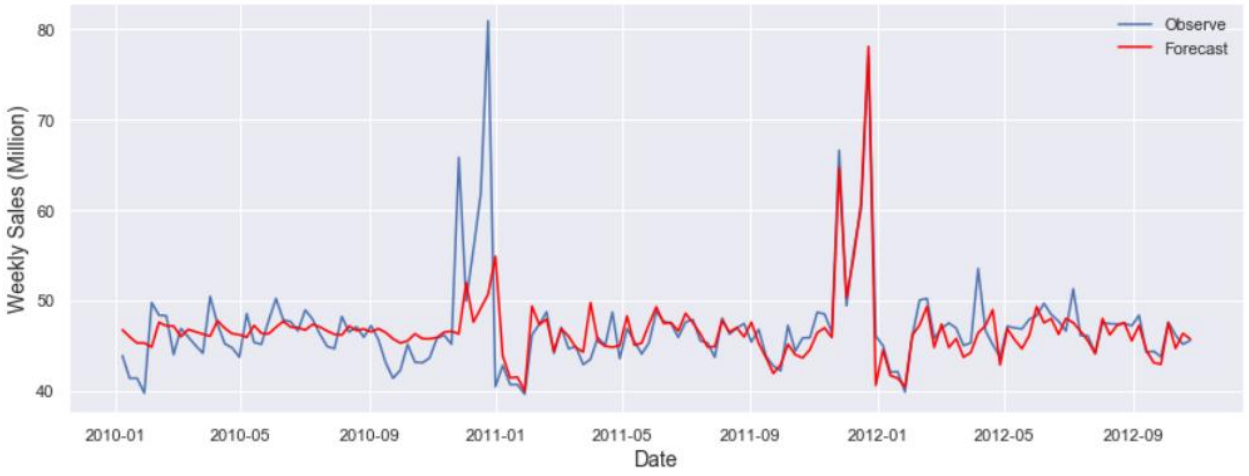
The diagnostic plots below show that the residuals fluctuate around the mean which is closed to zero. The Normal Q-Q plot shows that there are extreme values in the top right corner and bottom left corner.



The table below show the metrics for SARIMAX model. The mean squared error is around $15.5 million dollars. The Mean Absolute Percentage Error (MAPE) of 0.039452 indicates that the model is about 96% accurate in predicting the next 52 observations.

mape	me	mae	mpe	mse	rmse	corr	minmax
0.039452	-0.32611	1.964519	-0.001927	15.520546	3.939612	0.691568	0.037862

The plot below shows the model's observed and predicted sales. As shown on the plot, the model's prediction is more accurate after January 2011.



Sales Forecast

According to the model's performance metrics, SARIMAX model performs better than ARMA(1, 2) model. The SARIMAX model has lower AIC score, lower mean squared error, and lower mean absolute percentage error than that of ARMA model. Therefore, SARIMAX model is selected in making sales forecast. The plot below show sales forecast for the next 39 weeks.

Appendix

Appendix A – Jupyter Notebooks

- Data Wrangling
  https://github.com/nphan20181/walmart_sales/blob/master/walmart_data_wrangling.ipynb
- Exploratory Data Analysis
  https://github.com/nphan20181/walmart_sales/blob/master/walmart_eda1.ipynb
- Time Series analysis
  https://github.com/nphan20181/walmart_sales/blob/master/walmart_eda2.ipynb
- Time Series Models
  https://github.com/nphan20181/walmart_sales/blob/master/03_walmart_models.ipynb