

Analysis of Walmart Sales & Sales Forecast

NGOC PHAN, M.S. BUSINESS ANALYTICS

EMAIL: NPHAN20181@GMAIL.COM

GITHUB: [HTTPS://GITHUB.COM/NPHAN20181/WALMART_SALES](https://github.com/NPHAN20181/WALMART_SALES)

Agenda



PROJECT
BACKGROUND



MISSION
STATEMENT



DATASET



DATA
WRANGLING



EXPLORATORY
DATA ANALYSIS



TIME SERIES
ANALYSIS



FORECAST
MODELS



CONCLUSION

Special Buy \$148

onn. 50" Class 4K Roku Smart TV
 • 49.5" diagonal screen size
 • Model #100005396, 100005843, 100007147

onn. Roku
 NETFLIX HBO GO VUDU

PHILIPS
 androidtv
 NETFLIX HBO GO VUDU

Walmart

65" | 4K

Black Friday

Online Wed. 11/27 at 10pm ET
 In stores Thanksgiving 11/28 at 6pm

Special Buy \$98

onn. 40" Class 1080p Roku Smart TV
 • 39.5" diagonal screen size
 • Model #100005842, 100005395

onn. Roku
 NETFLIX HBO GO VUDU

Special Buy \$278

Philips 65" Class 4K Android Smart TV
 • 64.5" diagonal screen size
 • Model #65PFL5504/F7

Get up to **\$700** in eGift Cards

Walmart eGift Card

with qualified activation & trade-in on select Apple & Samsung smartphones.
 See page 9 for details.

Save \$80 \$249

iPad 7th Gen 32GB
 • 10.2" Retina display
 • Smart Keyboard & Apple Pencil compatible

Save \$70 \$129

Apple Watch Series 3 (GPS)
 • 38mm
 • \$199

Save \$15 \$129

Shop our **BUY NOW** selection

Project Background

- Walmart
 - an American multinational retail corporation
 - operates a chain of
 - hypermarkets
 - discount department stores
 - grocery stores
 - 45 stores across the U.S.
 - promotional markdown events
 - Super Bowl
 - Labor Day
 - Thanksgiving
 - Christmas

Mission Statement

- Assist Walmart's management team in the decision-making process by:
 - Performing exploratory data analysis and time series analysis of Walmart's sales data
 - Identifying the factors that impact sales
 - Developing machine learning algorithms to forecast sales

Dataset

- Collected on Kaggle at <https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data>
- Historical sales data
 - 45 Walmart stores in the United States
 - From 2/5/2010 to 11/1/2012
- 3 csv files
 - stores: 45 records
 - Columns: store, type, size
 - sales: 421,570 records
 - Columns: store, dept, date, weekly sales, isHoliday
 - features: 8,190 records
 - Columns: store, date, temp, fuel price, markdown 1-5, CPI, unemployment, isHoliday

Data Wrangling

MISSING VALUES | NEW COLUMNS | OUTLIERS

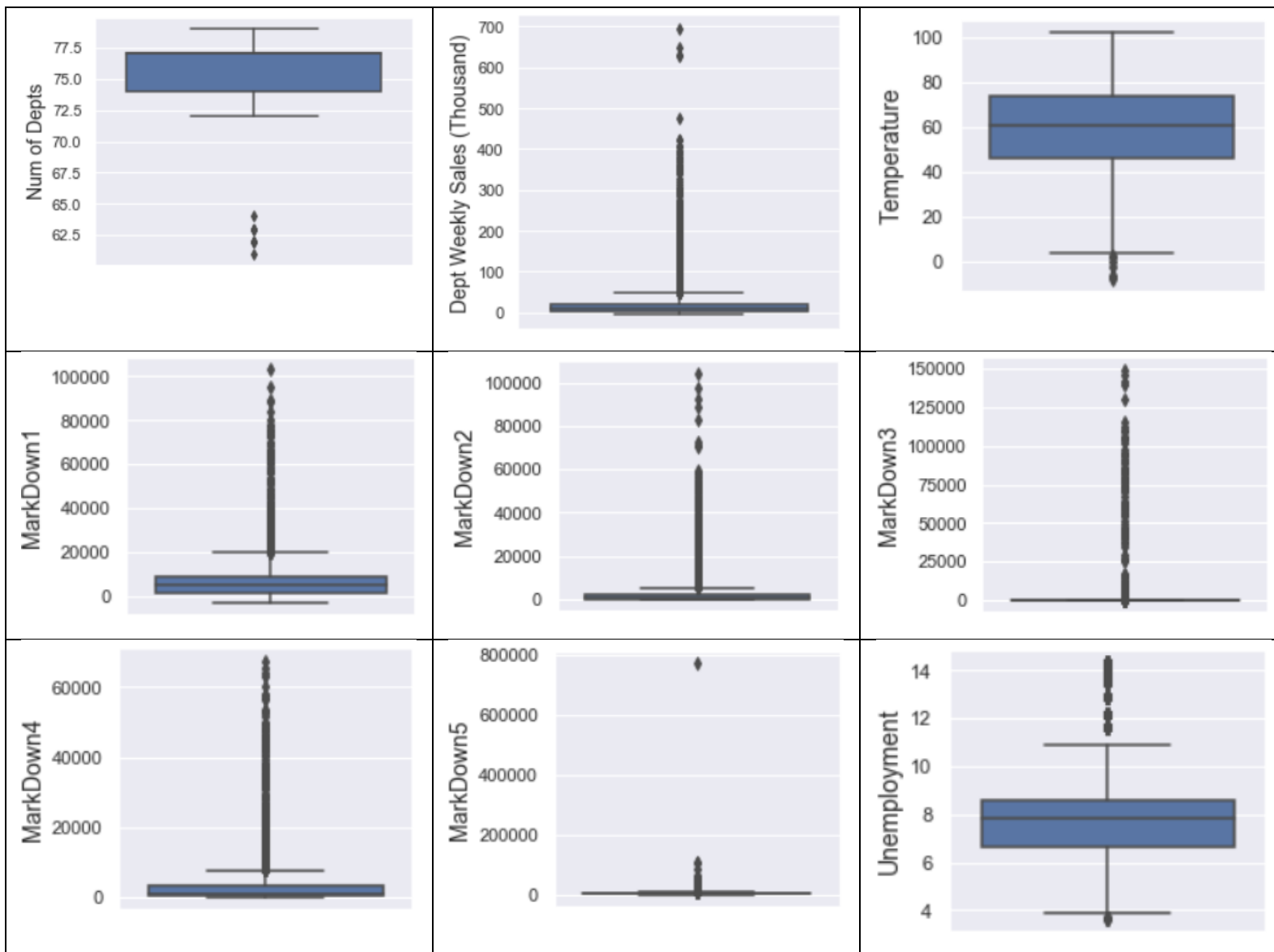
Missing Values

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

- features
 - Markdown 1-5
 - CPI
 - Unemployment

New Columns

- Num of Depts = counts of number of departments for each store
- Dept Weekly Sales (Thousand) = $\text{Weekly Sales} / 1,000$
- Avg Yearly Sales (Million) = average yearly sales by store
- Markdown = sum of Markdown1-5
- Avg Yearly Markdown (Thousand) = average yearly markdown by store
- Year = year extracted from Date
- Quarter = quarter extracted from Date
- Month = month extracted from Date
- Week = week of year extracted from Date



Outliers

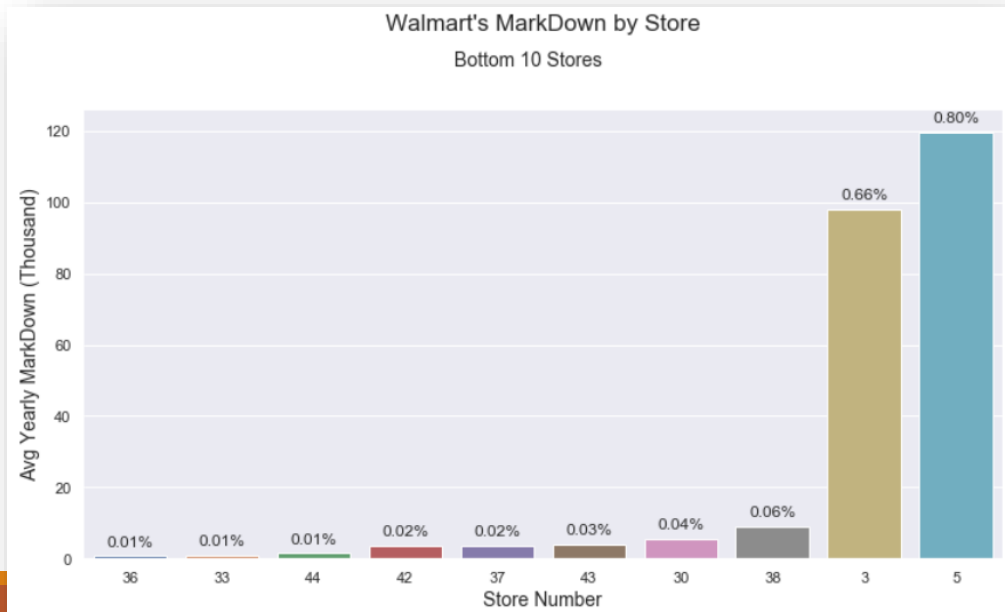
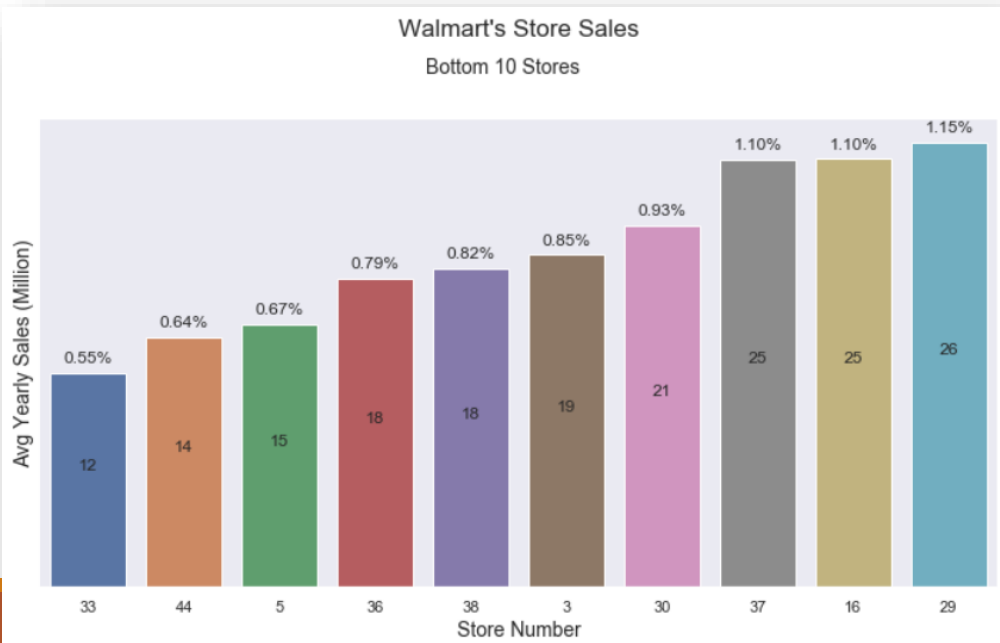
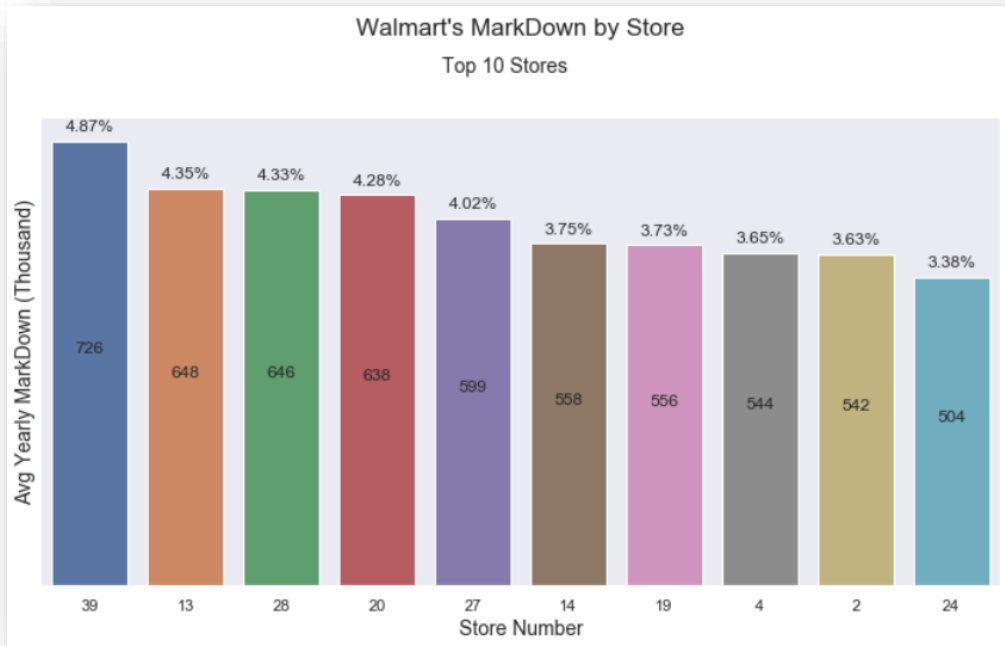
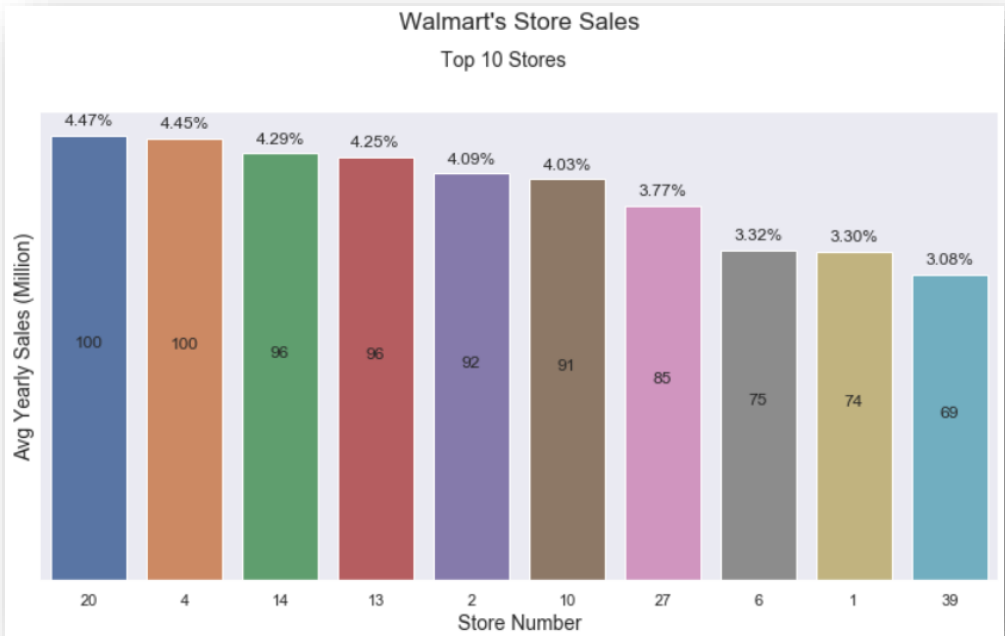
Exploratory Data Analysis

	Store_Sales_2010 (Million)	Store_Sales_2011 (Million)	Store_Sales_2012 (Million)
count	45.000000	45.000000	45.000000
mean	50.864136	54.404445	44.447397
std	26.783837	28.592598	23.019093
min	12.766834	12.957837	11.435551
25%	25.568078	29.117303	24.827531
50%	48.370384	50.360182	41.739164
75%	66.890648	74.169226	59.212433
max	105.462242	111.092293	92.771189

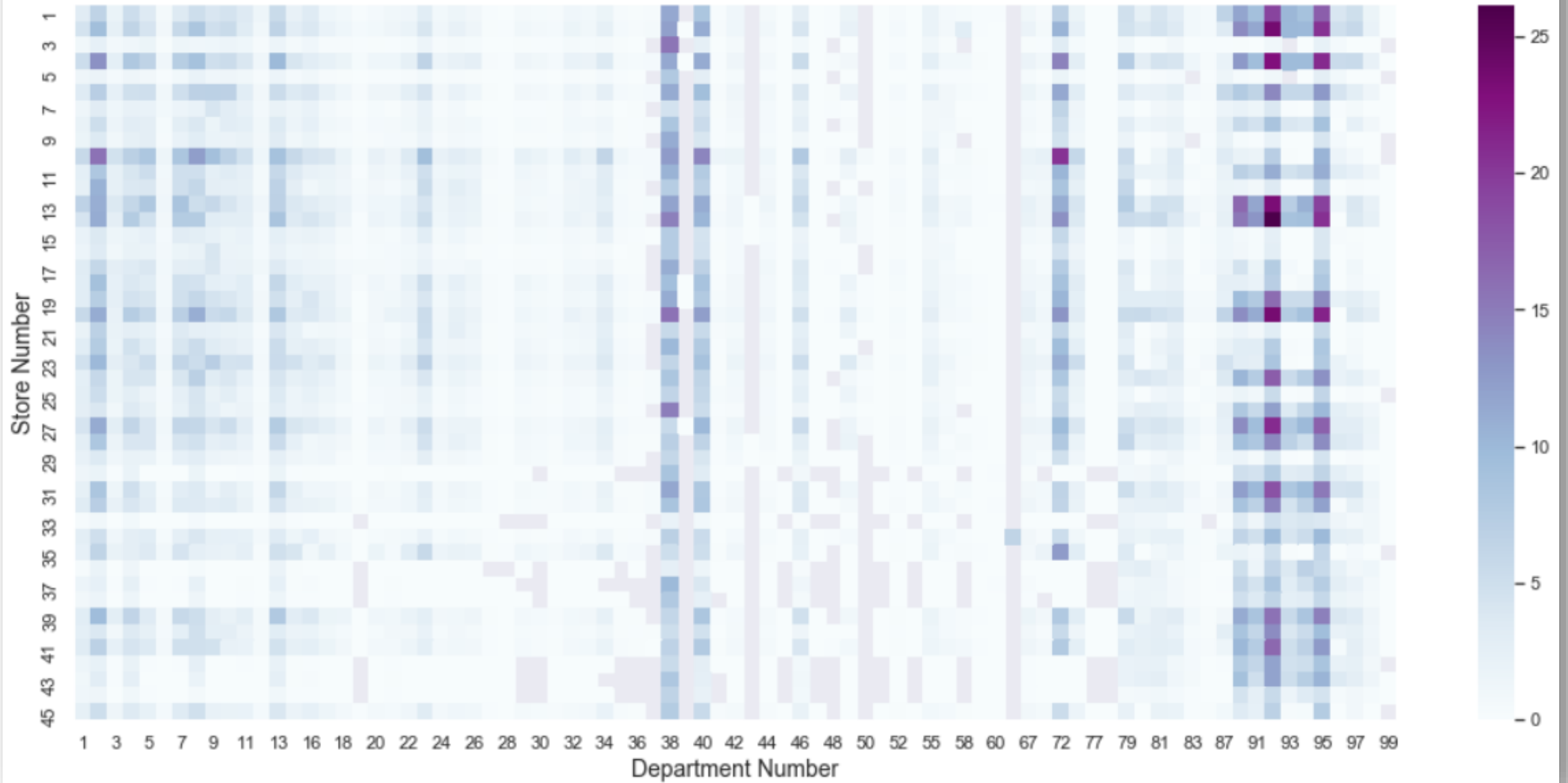
Store Weekly Sales (Thousand)	
count	6435.000000
mean	1046.964878
std	564.366622
min	209.986250
25%	553.350105
50%	960.746040
75%	1420.158660
max	3818.686450

Dept Weekly Sales (Thousand)	
count	421570.000000
mean	15.981258
std	22.711184
min	-4.988940
25%	2.079650
50%	7.612030
75%	20.205853
max	693.099360

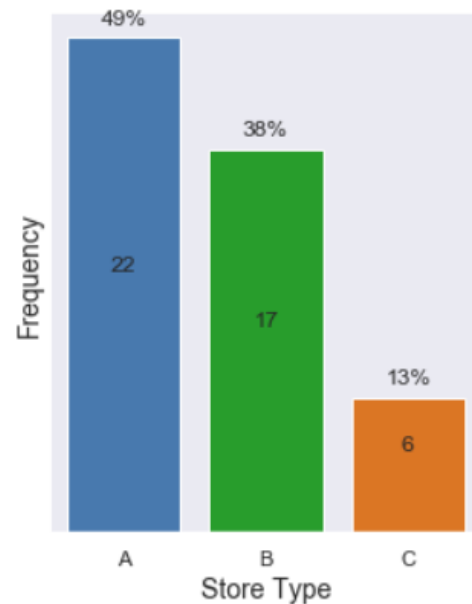




Walmart's Store Sales by Department (Million USD)

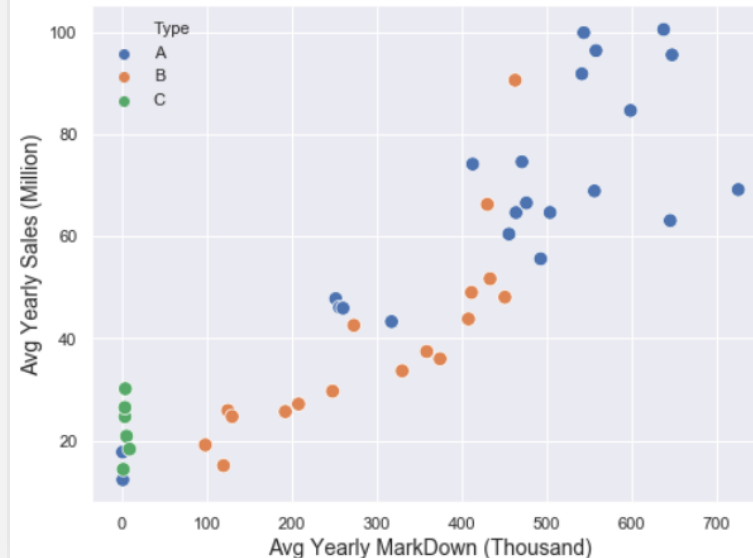


Distribution of Walmart's Store Types



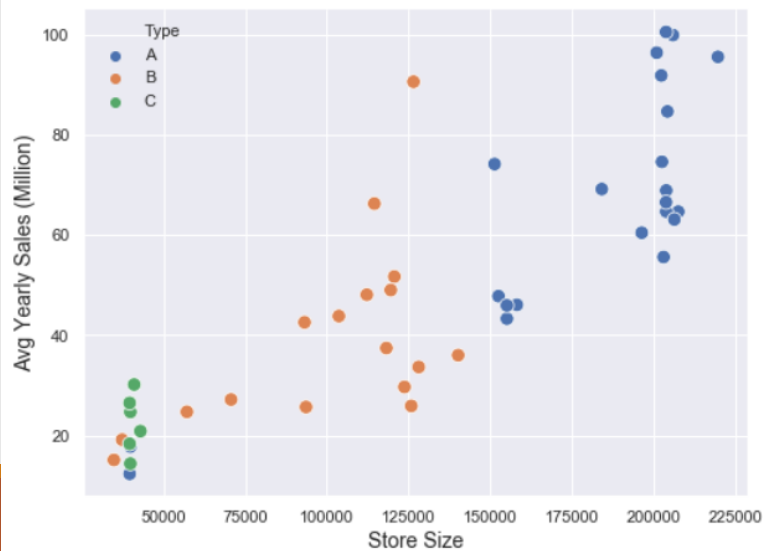
Walmart Sales

Avg Yearly MarkDown (Thousand) vs. Avg Yearly Sales (Million)
Correlation: 0.8778



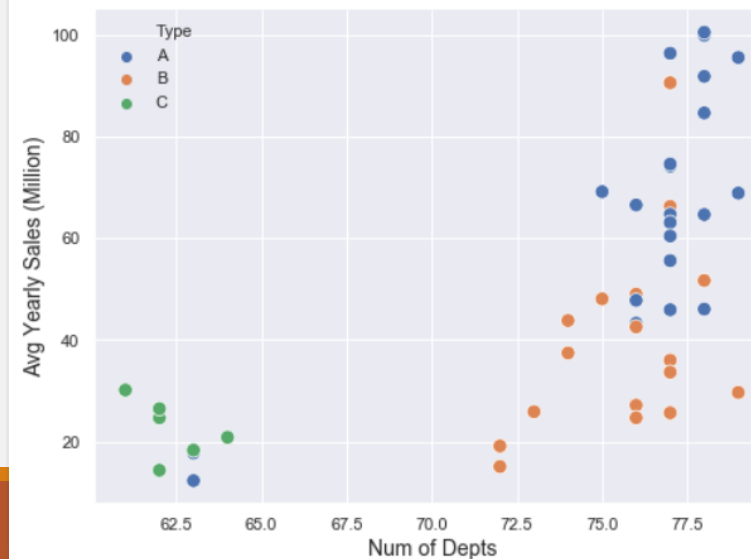
Walmart Sales

Size vs. Avg Yearly Sales (Million)
Correlation: 0.8462



Walmart Sales

Num of Depts vs. Avg Yearly Sales (Million)
Correlation: 0.6385



Correlation



Weak
Correlation

Time Series Analysis

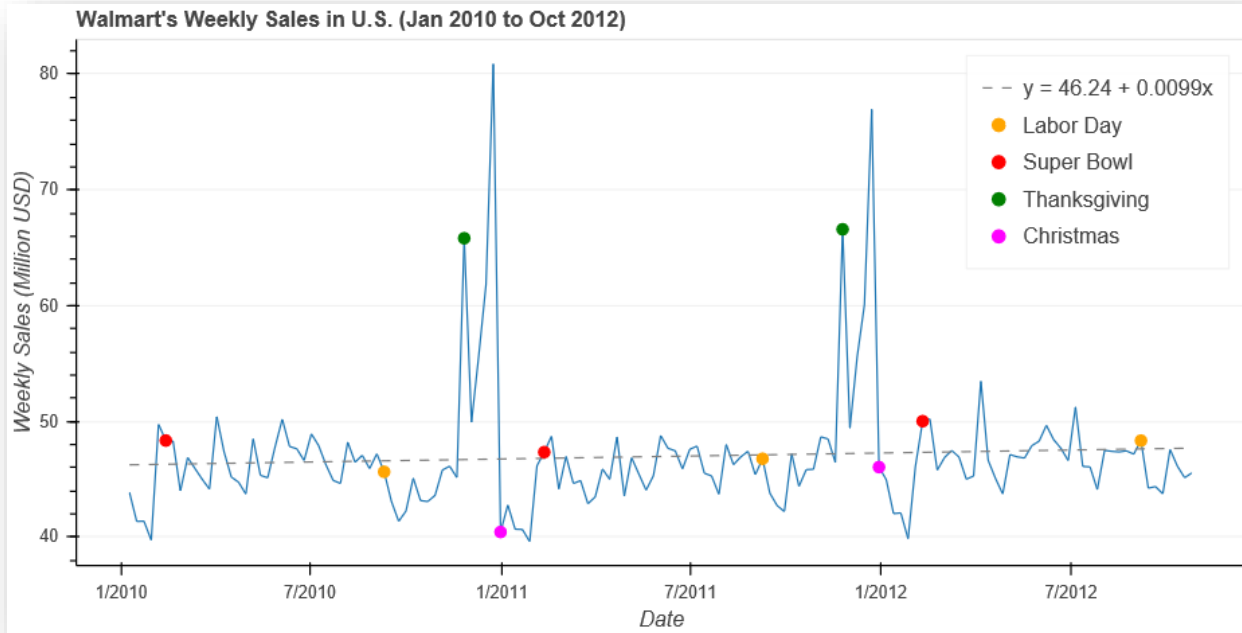
ORIGINAL TIME
SERIES

GENERAL TREND

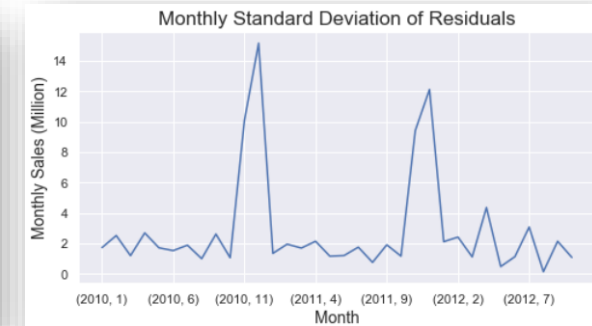
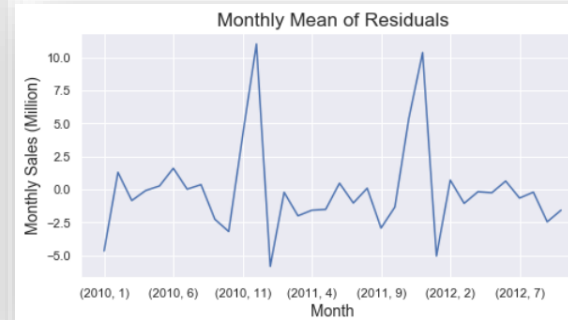
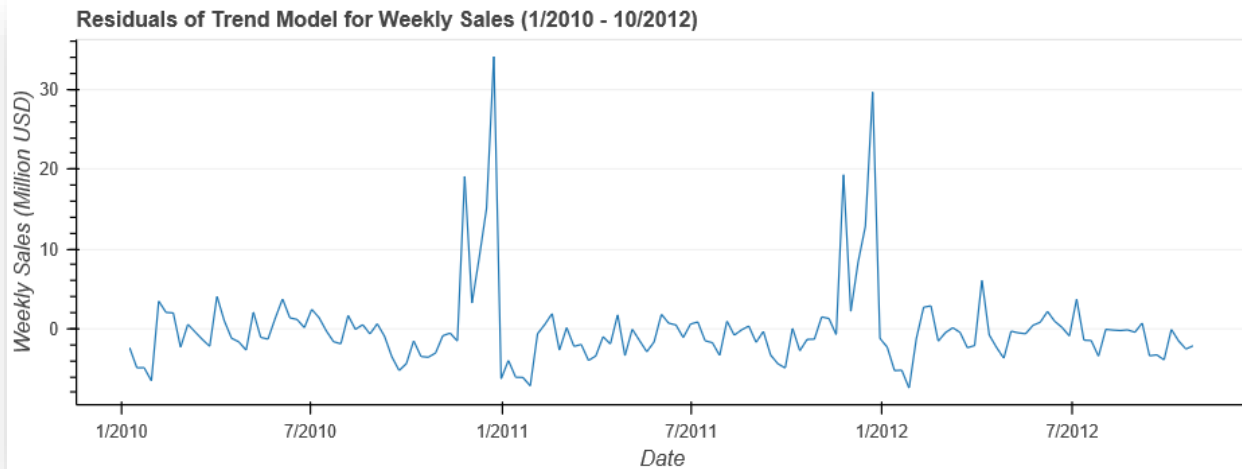
SEASONALITY

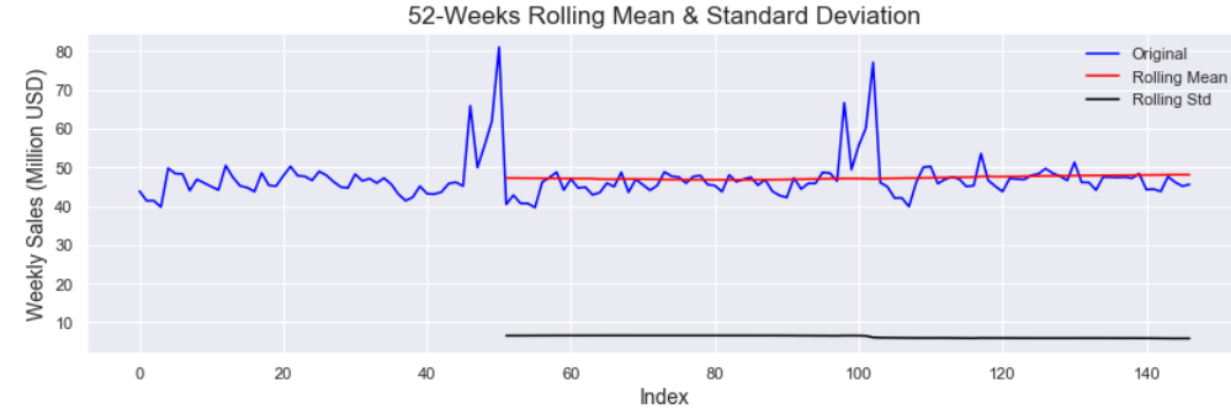
STATIONARITY

Original Time Series & General Trend (y)



Seasonality





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

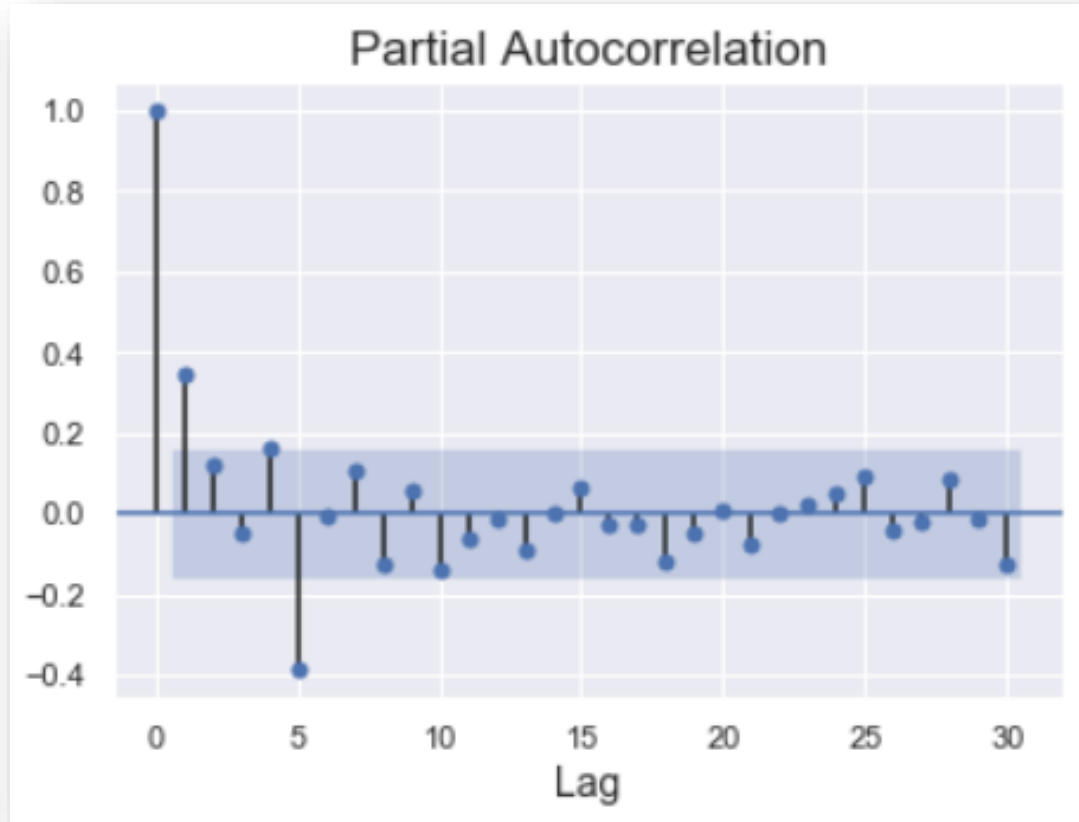
Stationarity

Model Development

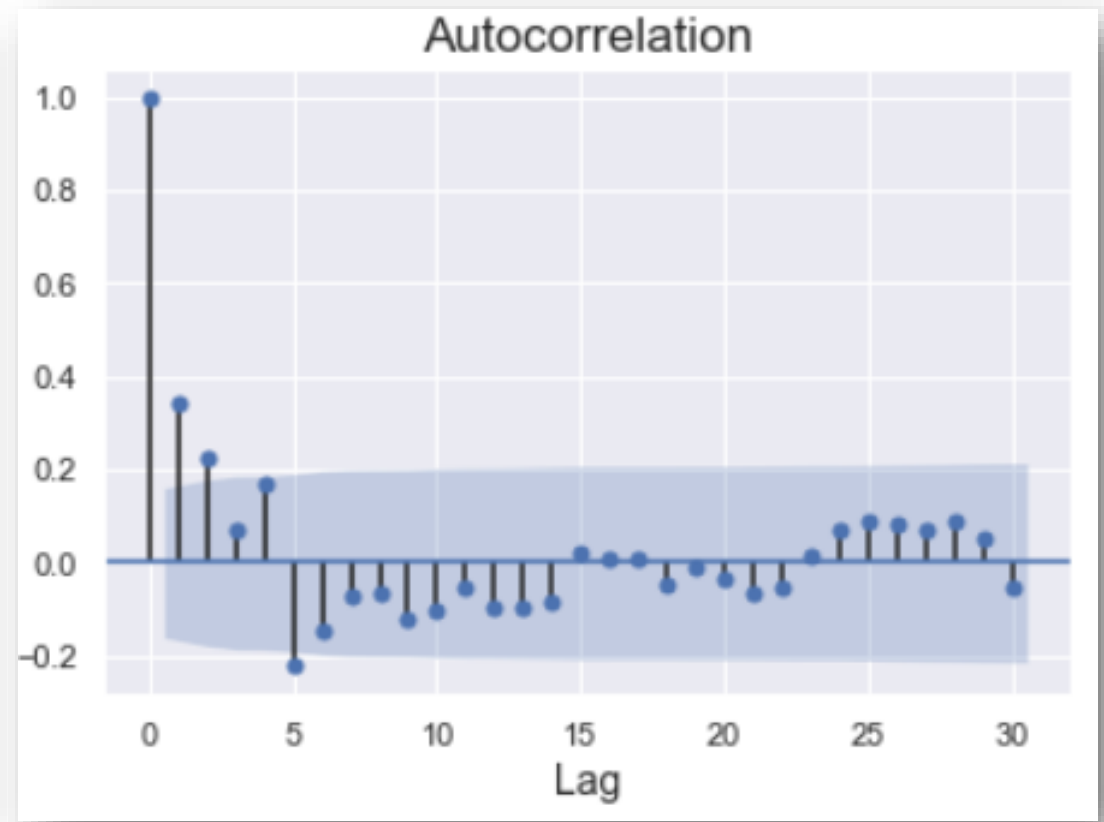
AUTOREGRESSIVE TERM | MOVING AVERAGE TERM

ARIMA | SARIMAX

Autoregressive Term



Moving Average Term



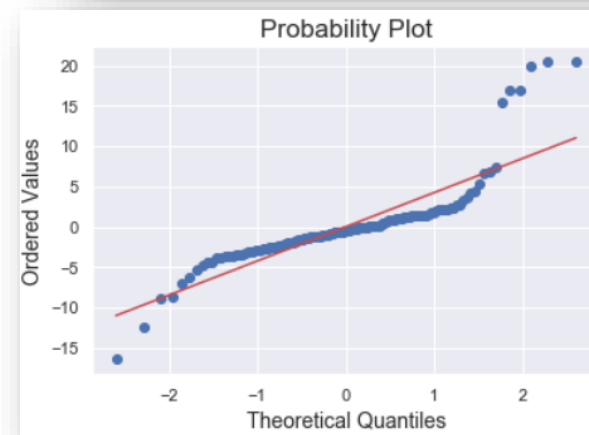
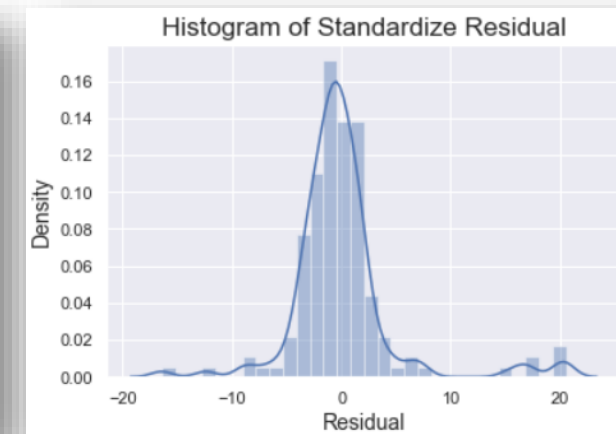
ARIMA Model

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



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

SARIMAX Model

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

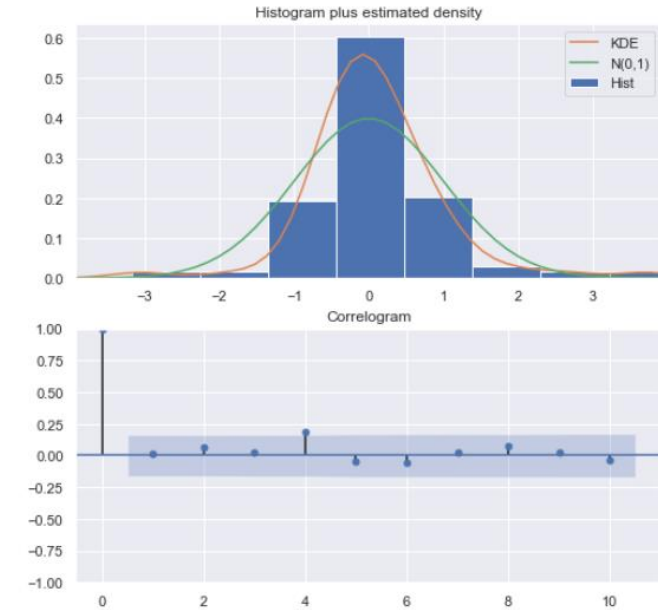
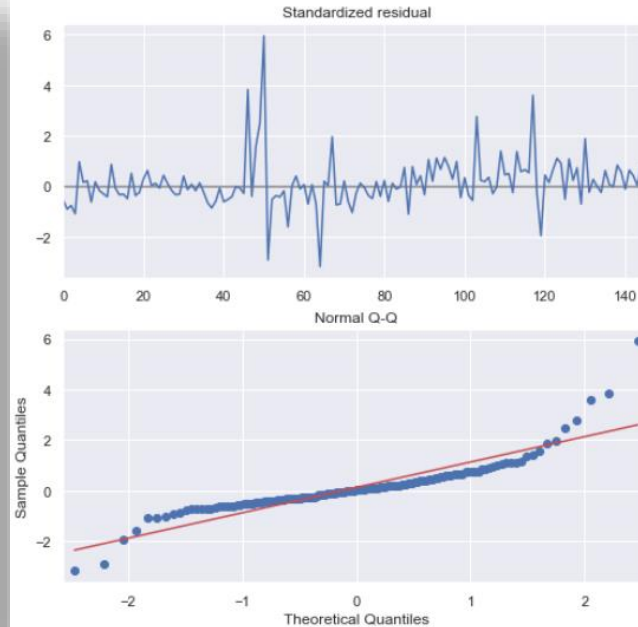
	coef	std err	z	P> z	[0.025	0.975]
intercept	2.6581	0.431	6.161	0.000	1.813	3.504
ar.L1	0.2734	0.052	5.254	0.000	0.171	0.375
ar.S.L52	0.9217	0.009	98.294	0.000	0.903	0.940
sigma2	3.9129	0.379	10.330	0.000	3.170	4.655

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



Evaluation Metrics

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

Sales Forecast

Sales Forecast from ARIMA Model



Sales Forecast from SARIMAX Model



Future Sales Forecast from SARIMAX Model



Conclusion

- Promotional markdown events before holidays seem to increase sales except for Christmas
- Sales seems to be highest
 - During the week of Thanksgiving
 - 2-3 weeks after Thanksgiving
- Stores with high sales: type A stores
 - Big size
 - Big number of departments
 - High markdown values
- Stores with low sales: type C stores
 - Small size
 - Small number of departments
 - Low markdown values

Thank You!