Retail Analysis with Walmart Data

DESCRIPTION

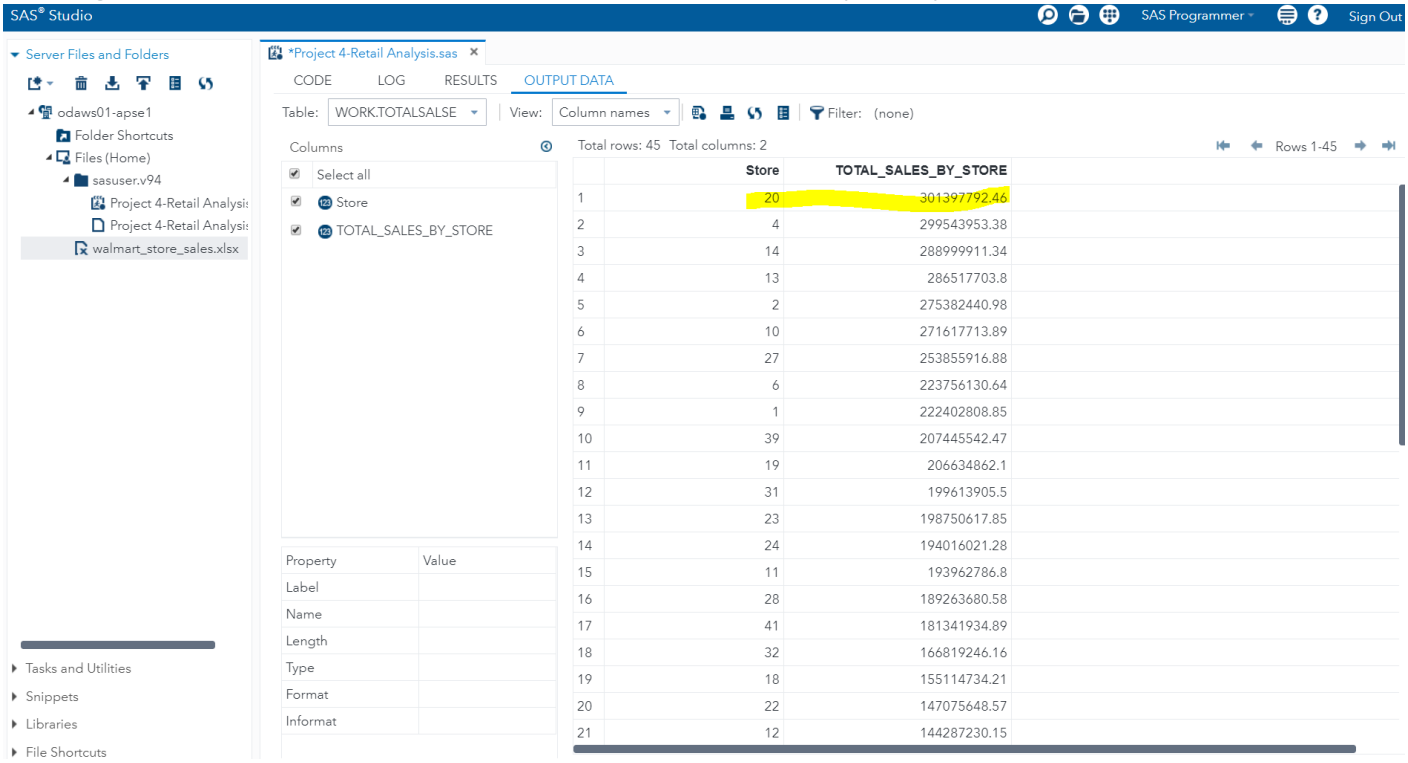
One of the leading retail stores in the US, Walmart, would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for 45 stores of Walmart. The business is facing a challenge due to unforeseen demands and runs out of stock some times, due to the inappropriate machine learning algorithm. An ideal ML algorithm will predict demand at different points of time covering seasonality and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of all, which are the Super Bowl, Labour Day, Thank

<u>Basic Statistics tasks</u>

- Which store has maximum sales- Store 20

We will give an conclusion that store 20 have maximum sales when sum up weekly sales in the dataset.



	Store	TOTAL_SALES_BY_STORE
1	20	301397792.46
2	4	299543953.38
3	14	288999911.34
4	13	286517703.8
5	2	275382440.98
6	10	271617713.89
7	27	253855916.88
8	6	223756130.64
9	1	222402808.85
10	39	207445542.47
11	19	206634862.1
12	31	199613905.5
13	23	198750617.85
14	24	194016021.28
15	11	193962786.8
16	28	189263680.58
17	41	181341934.89
18	32	166819246.16
19	18	155114734.21
20	22	147075648.57
21	12	144287230.15

- Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation

Answer is Store 14 has maximum standard deviation.

Store has maximum standard deviation

Obs	Store	STDBYSALES	MEANBYSALES
1	14	317569.94948	2020978.401
2	10	302262.0625	1899424.5727
3	20	275900.56274	2107676.8703
4	4	266201.4423	2094712.9607
5	13	265506.99578	2003620.3063
6	23	249788.03807	1389864.4605
7	27	239930.13569	1775216.202
8	2	237683.69468	1925751.3355
9	39	217466.45483	1450668.1292
10	6	212525.85586	1564728.1863
11	35	211243.45779	919724.97958
12	19	191722.63873	1444999.0357
13	41	187907.16277	1268125.4188
14	28	181758.96754	1323522.2418
15	18	176641.51084	1084718.421
16	24	167745.67757	1356755.3936
17	11	165833.88786	1356383.1245
18	22	161251.35063	1028501.039
19	1	155980.76776	1555264.3976
20	12	139166.87188	1009001.6094
21	32	138017.25209	1166568.155
22	45	130168.52664	785981.40853
23	21	128752.81285	756069.08336
24	31	125855.94293	1395901.4371
25	15	120538.65204	623312.47497
26	40	119002.11286	964128.04049
27	25	112976.7886	706721.53266
28	7	112585.46922	570617.30867
29	17	112162.93609	893581.39042
30	26	110431.28814	1002911.8447
31	8	106280.82988	908749.51839
32	34	104630.16468	966781.55944

Coefficient of mean to standard deviation

The MEANS Procedure

Analysis Variable : Weekly_Sales		
Store	N Obs	Coeff of Variation
1	143	10.0292123
2	143	12.3423876
3	143	11.5021407
4	143	12.7082539
5	143	11.8668441
6	143	13.5822859
7	143	19.7304687
8	143	11.6952832
9	143	12.6895468
10	143	15.9133491
11	143	12.2261834
12	143	13.7925322
13	143	13.2513628
14	143	15.7136736
15	143	19.3383988
16	143	16.5180655
17	143	12.5520671
18	143	16.2845497
19	143	13.2680115
20	143	13.0902686
21	143	17.0292392
22	143	15.6782876
23	143	17.9721149
24	143	12.3637377
25	143	15.9860402
26	143	11.0110663
27	143	13.5155445
28	143	13.7329742
29	143	18.3742467
30	143	5.2008039
31	143	9.0161053
32	143	11.8310492
33	143	9.2868353

- Which store/s has good quarterly growth rate in Q3’2012

In order to address this, we need to filter data set by year = 2012 and Qtr=3 (July Aug September)

Then follow by transform data to SAS timeseries dataset

Obs	Store	Date	GROWTH_RATE
1	28	2012:3	14.82%
2	17	2012:3	10.36%
3	42	2012:3	7.15%
4	44	2012:3	5.98%
5	33	2012:3	5.41%
6	38	2012:3	4.73%
7	37	2012:3	4.12%
8	39	2012:3	3.46%
9	14	2012:3	2.73%
10	8	2012:3	2.12%
11	30	2012:3	0.96%
12	29	2012:3	0.89%
13	19	2012:3	0.70%
14	10	2012:3	0.41%
15	20	2012:3	0.13%
16	34	2012:3	(0.32%)
17	43	2012:3	(0.42%)
18	40	2012:3	(0.66%)
19	32	2012:3	(0.95%)
20	41	2012:3	(1.78%)
21	1	2012:3	(2.93%)
22	11	2012:3	(3.02%)
23	13	2012:3	(3.59%)
24	4	2012:3	(3.69%)
25	5	2012:3	(3.81%)
26	9	2012:3	(3.82%)
27	25	2012:3	(3.87%)
28	31	2012:3	(4.05%)
29	35	2012:3	(4.67%)
30	18	2012:3	(4.71%)
31	12	2012:3	(5.13%)
32	22	2012:3	(5.23%)

- Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together

First, we will filter out holiday sales data with holiday flag=1

Second, we need to calculate the mean of holiday sales by store.

Third, base on the mean value, we will categorize those stores with higher sales when holiday sales is bigger than mean value.

The MEANS Procedure

Analysis Variable : Weekly_Sales				
N	Mean	Std Dev	Minimum	Maximum
5985	1041256.38	558957.44	209986.25	3818686.45

Higher Sales during Holidays

Obs	Store	Weekly_Sales	Date	HOLIDAY
1	1	1641957.44	12/02/2010	1
2	1	1507460.69	10/09/2010	1
3	1	1955624.11	26/11/2010	1
4	1	1367320.01	31/12/2010	1
5	1	1649614.93	11/02/2011	1
6	1	1540471.24	09/09/2011	1
7	1	2033320.66	25/11/2011	1
8	1	1497462.72	30/12/2011	1
9	1	1802477.43	10/02/2012	1
10	1	1661767.33	07/09/2012	1
11	2	2137809.5	12/02/2010	1
12	2	1839128.83	10/09/2010	1
13	2	2658725.29	26/11/2010	1
14	2	1750434.55	31/12/2010	1
15	2	2168041.61	11/02/2011	1
16	2	1748000.65	09/09/2011	1
17	2	2614202.3	25/11/2011	1
18	2	1874226.52	30/12/2011	1
19	2	2103322.68	10/02/2012	1
20	2	1898777.07	07/09/2012	1
21	4	2188307.39	12/02/2010	1
22	4	1865820.81	10/09/2010	1
23	4	2789469.45	26/11/2010	1
24	4	1794868.74	31/12/2010	1
25	4	2187847.29	11/02/2011	1
26	4	2093139.01	09/09/2011	1
27	4	3004702.33	25/11/2011	1
28	4	2007105.86	30/12/2011	1
29	4	2374660.64	10/02/2012	1

- Provide a monthly and semester view of sales in units and give insights

First, we need to transform dataset to timeseries dataset.

Second, set interval as month.

SAS will automatically sum up sales by month.

Obs	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	1	FEB2010	$6,307,344.10	1	167.38	10.195	844.9473138	32.424
2	1	MAR2010	$5,871,293.98	0	210.32	10.744	844.9644632	32.424
3	1	APR2010	$7,422,801.92	0	326.7	13.872	1052.7606757	39.04
4	1	MAY2010	$5,929,938.64	0	304.21	11.274	842.1912485	31.232
5	1	JUN2010	$6,084,081.46	0	329.57	10.663	845.4249474	31.232
6	1	JUL2010	$7,244,483.04	0	409.74	13.182	1056.0373802	38.935
7	1	AUG2010	$6,075,952.95	0	346.03	10.602	846.3147521	31.148
8	1	SEP2010	$5,829,793.92	1	322.95	10.348	846.1461223	31.148
9	1	OCT2010	$7,150,641.75	0	342.5	13.397	1059.002625	39.19
10	1	NOV2010	$6,485,547.06	1	234.28	10.923	847.5984449	31.352
11	1	DEC2010	$8,876,953.18	1	246.2	14.249	1057.2885123	39.19
12	1	JAN2011	$5,480,050.97	0	171.54	11.985	846.8864448	30.968
13	1	FEB2011	$6,399,887.57	1	198.92	12.121	852.2870801	30.968
14	1	MAR2011	$6,307,375.48	0	245.87	13.708	856.8970394	30.968
15	1	APR2011	$7,689,123.60	0	343.3	18.506	1076.2912761	38.41
16	1	MAY2011	$6,128,431.80	0	285.6	15.498	862.9977641	30.728
17	1	JUN2011	$6,194,971.74	0	336.12	14.578	860.406336	30.728
18	1	JUL2011	$7,227,654.31	0	432.52	17.912	1076.7281475	39.81
19	1	AUG2011	$6,144,985.73	0	360.31	14.399	862.5529423	31.848
20	1	SEP2011	$7,379,542.34	1	399.26	17.427	1080.7856706	39.81
21	1	OCT2011	$6,072,327.75	0	271.33	13.284	869.593297	31.464
22	1	NOV2011	$6,864,972.83	1	236.48	13.173	872.5235968	31.464
23	1	DEC2011	$9,032,594.71	1	236.98	15.73	1095.7497436	39.33
24	1	JAN2012	$5,723,690.52	0	205.91	12.976	879.671326	29.392
25	1	FEB2012	$6,798,074.91	1	207.14	13.834	881.4998542	29.392
26	1	MAR2012	$8,201,997.40	0	318	18.665	1105.7664718	36.74
27	1	APR2012	$6,511,214.82	0	273.49	15.473	886.1278322	28.572
28	1	MAY2012	$6,446,962.46	0	296.87	14.628	886.8850812	28.572
29	1	JUN2012	$8,020,582.84	0	398.87	16.978	1108.905761	35.715
30	1	JUL2012	$6,233,946.67	0	321.77	13.201	887.6819589	27.632
31	1	AUG2012	$7,897,619.59	0	414.16	17.74	1110.4241342	34.54
32	1	SEP2012	$6,122,381.52	1	304.88	14.834	890.7845311	27.632
33	1	OCT2012	$6,245,587.29	0	268.67	14.318	893.4327483	26.292
34	2	FEB2010	$8,264,347.77	1	164.47	10.195	843.5717166	33.296
35	2	MAR2010	$7,677,765.60	0	210.51	10.744	843.5942463	33.296

From below corrlation table, weekly sales is relate to CPI, Fuel price, unemployment features.

The CORR Procedure

8 Variables:	Store Date Weekly_Sales Holiday_Flag Temperature Fuel_Price CPI Unemployment

Simple Statistics

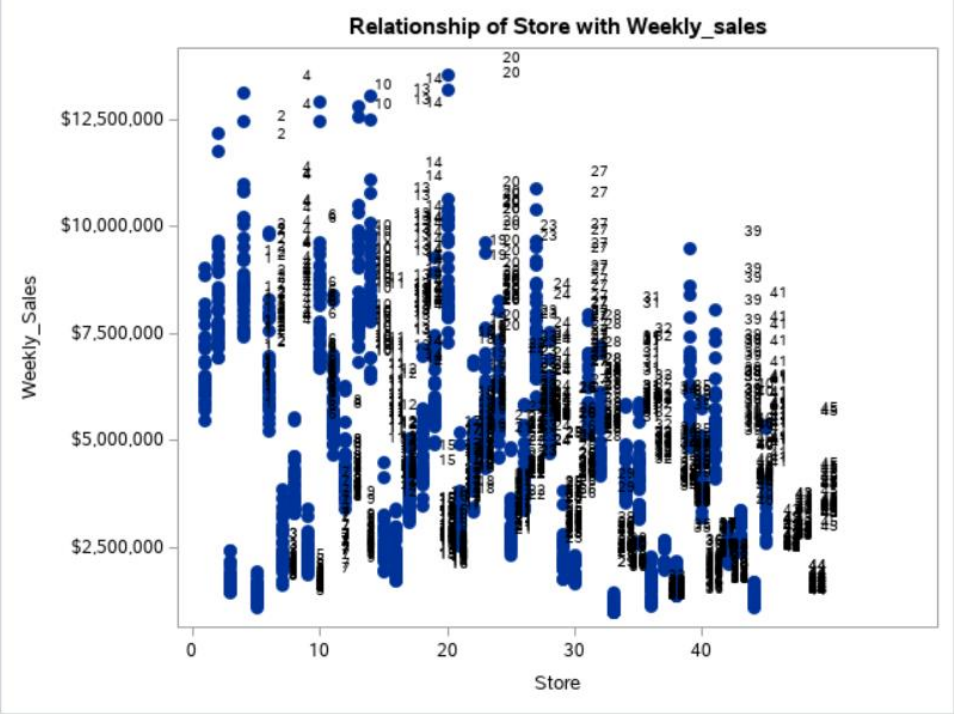
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Store	1485	23.00000	12.99155	34155	1.00000	45.00000
Date	1485	18779	290.04230	27887535	18294	19267
Weekly_Sales	1485	4536848	2480845	6737218987	978511	13553792
Holiday_Flag	1485	0.30303	0.45972	450.00000	0	1.00000
Temperature	1485	262.87639	86.27854	390371	42.94000	483.43000
Fuel_Price	1485	14.55396	2.54783	21613	10.07600	21.03300
CPI	1485	743.50637	189.74914	1104107	504.31539	1129
Unemployment	1485	34.66299	9.03514	51475	15.51600	71.56500

Pearson Correlation Coefficients, N = 1485
Prob > |r| under H0: Rho=0

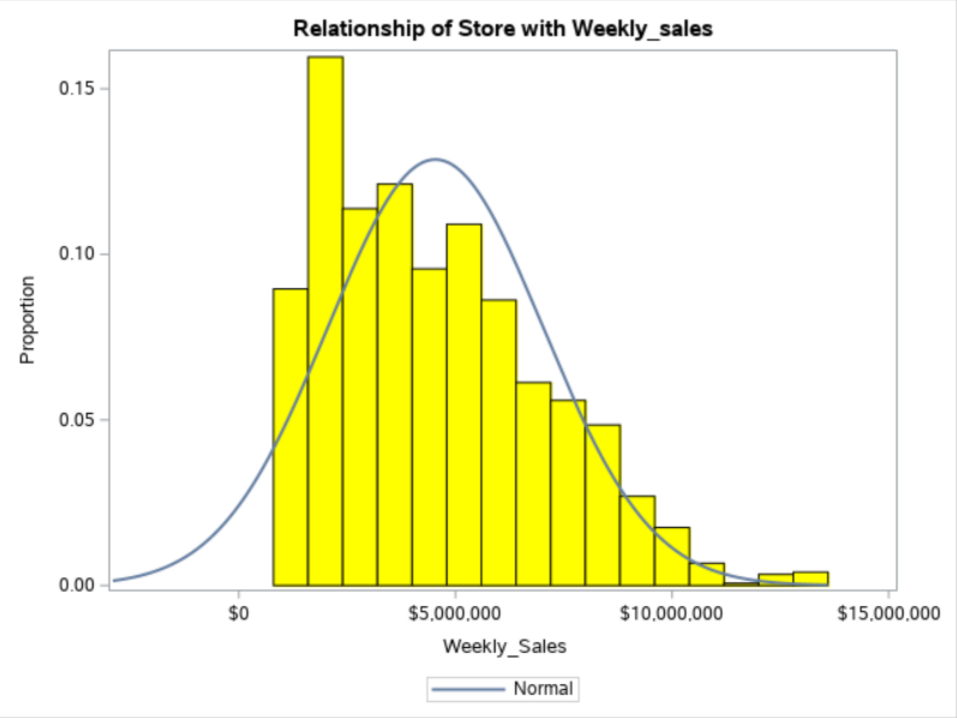
	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
Store	1.00000	0.00000 1.0000	-0.33065 <.0001	0.00000 1.0000	-0.02100 0.4188	0.04687 0.0710	-0.18834 <.0001	0.20116 <.0001
Date	0.00000 1.0000	1.00000	0.01263 0.6268	-0.03317 0.2014	0.14228 <.0001	0.61620 <.0001	0.08037 0.0019	-0.21141 <.0001
Weekly_Sales	-0.33065 <.0001	0.01263 0.6268	1.00000	0.05658 0.0292	0.03374 0.1938	0.15166 <.0001	0.03260 0.2093	0.00041 0.9875
Holiday_Flag	0.00000 1.0000	-0.03317 0.2014	0.05658 0.0292	1.00000	-0.32071 <.0001	-0.16177 <.0001	-0.02412 0.3531	0.00520 0.8413
Temperature	-0.02100 0.4188	0.14228 <.0001	0.03374 0.1938	-0.32071 <.0001	1.00000	0.37909 <.0001	0.33452 <.0001	0.27012 <.0001
Fuel_Price	0.04687 0.0710	0.61620 <.0001	0.15166 <.0001	-0.16177 <.0001	0.37909 <.0001	1.00000	0.14791 <.0001	0.24459 <.0001
CPI	-0.18834 <.0001	0.08037 0.0019	0.03260 0.2093	-0.02412 0.3531	0.33452 <.0001	0.14791 <.0001	1.00000	-0.06532 0.0118
Unemployment	0.20116 <.0001	-0.21141 <.0001	0.00041 0.9875	0.00520 0.8413	0.27012 <.0001	0.24459 <.0001	-0.06532 0.0118	1.00000

Lets look insight of the data

Store with Weekly Sales through scatter plot



Weekly sales distribution



Statistical Model
For Store 1 – Build prediction models to forecast demand

- Linear Regression – Utilize variables like date and restructure dates as 1 for 5 Feb 2010(starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.

When we implement Linear regression on monthly sales data, Pr>F = 0.0062 which means monthly salse is significantly relate to CPI, unemployment, and fuel price

However, CPI and unemployment have positive impact on the sales, Fuel price have negative impact.

The REG Procedure
Model: MODEL1
Dependent Variable: Weekly_Sales

Number of Observations Read	143
Number of Observations Used	143

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.936132E11	97871067831	4.30	0.0062
Error	139	3.161247E12	22742782618		
Corrected Total	142	3.45486E12			

Root MSE	150807	R-Square	0.0850
Dependent Mean	1555264	Adj R-Sq	0.0652
Coeff Var	9.69656		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3887096	1740276	-2.23	0.0271
CPI	1	21792	6785.27208	3.21	0.0016
Unemployment	1	124064	58779	2.11	0.0366
Fuel_Price	1	-64838	46841	-1.38	0.1685

- Time series forecasting model –
    - Hypothesize if the data is fit for time series analysis – check for white noise probability test
    - Make adjustments in historical data for events like holidays, if applicable
    - Build ARIMA model to forecast 6 months i.e., input utilize only till April 2012.

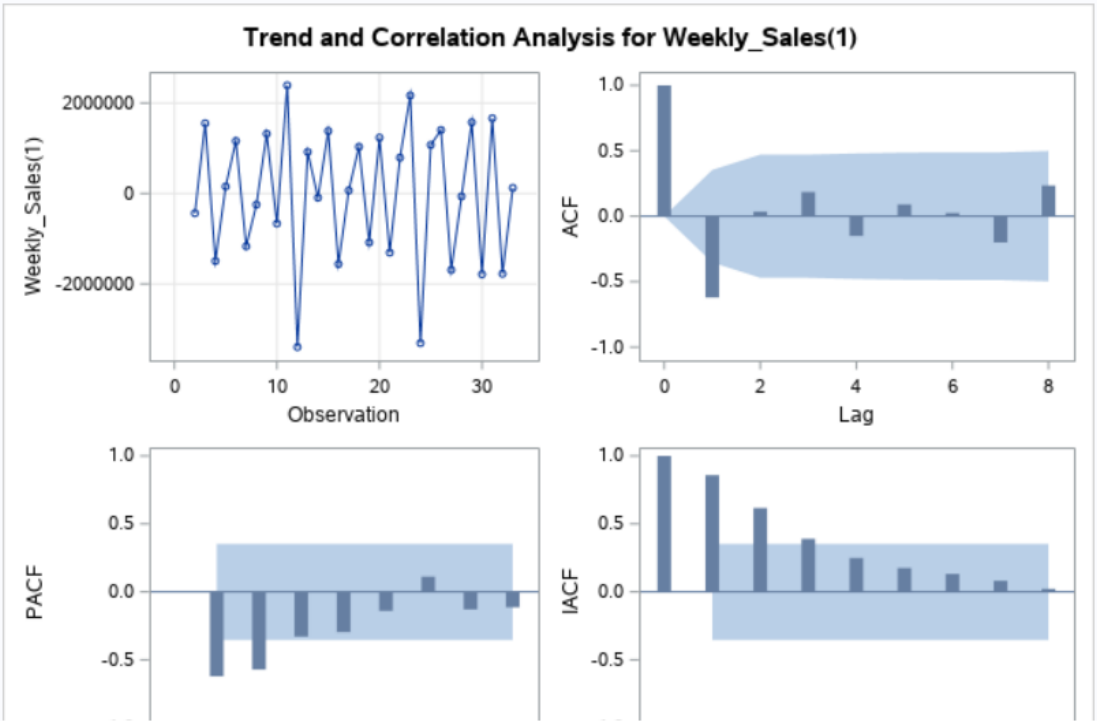
Predict next 6 months i.e., June to Oct 2010. Check for MAPE.

First, we still need to transform monthly sales data into timeseries format. And using SAS ARIMA to forecast 6 months data.

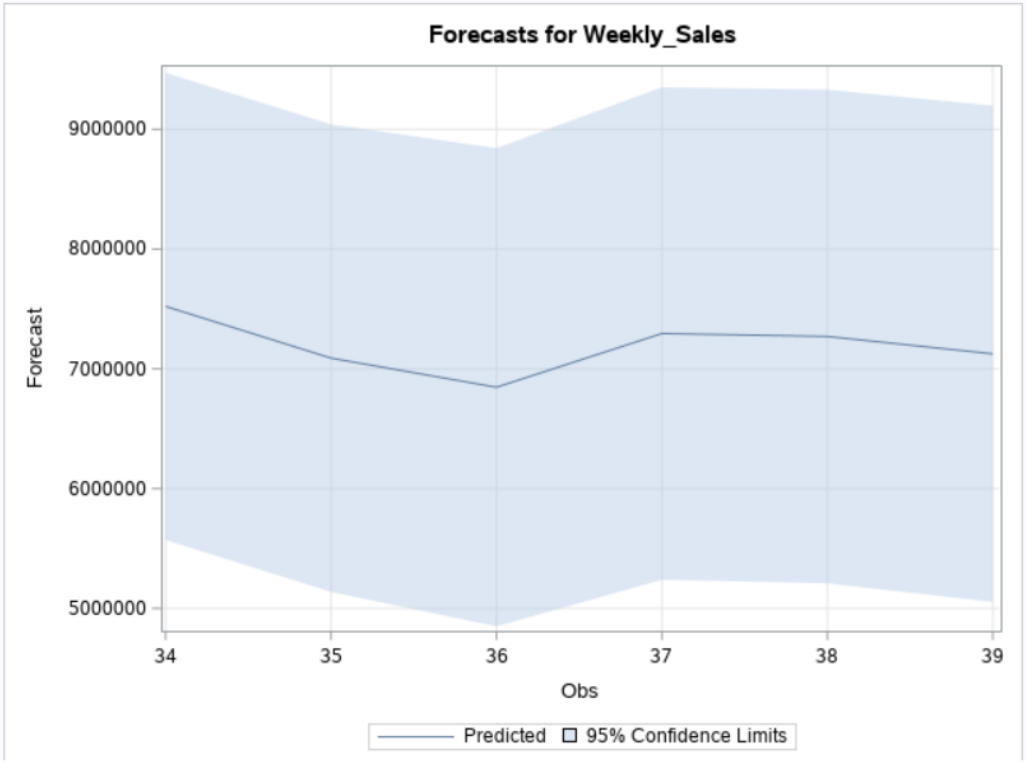
The ARIMA Procedure

Name of Variable = Weekly_Sales	
Period(s) of Differencing	1
Mean of Working Series	-1929.9
Standard Deviation	1495924
Number of Observations	32
Observation(s) eliminated by differencing	1

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	16.11	6	0.0132	-0.621	0.035	0.185	-0.151	0.090	0.024



Forecasts for variable Weekly_Sales				
Obs	Forecast	Std Error	95% Confidence Limits	
34	7521638.6	994721	5572021.8	9471255.4
35	7089643.6	995691	5138125.3	9041161.9
36	6846372.3	1018400	4850344.5	8842400.1
37	7295015.8	1048887	5239235.9	9350795.8
38	7270354.2	1051220	5210000.6	9330707.8
39	7125633.8	1056837	5054271.0	9196996.7

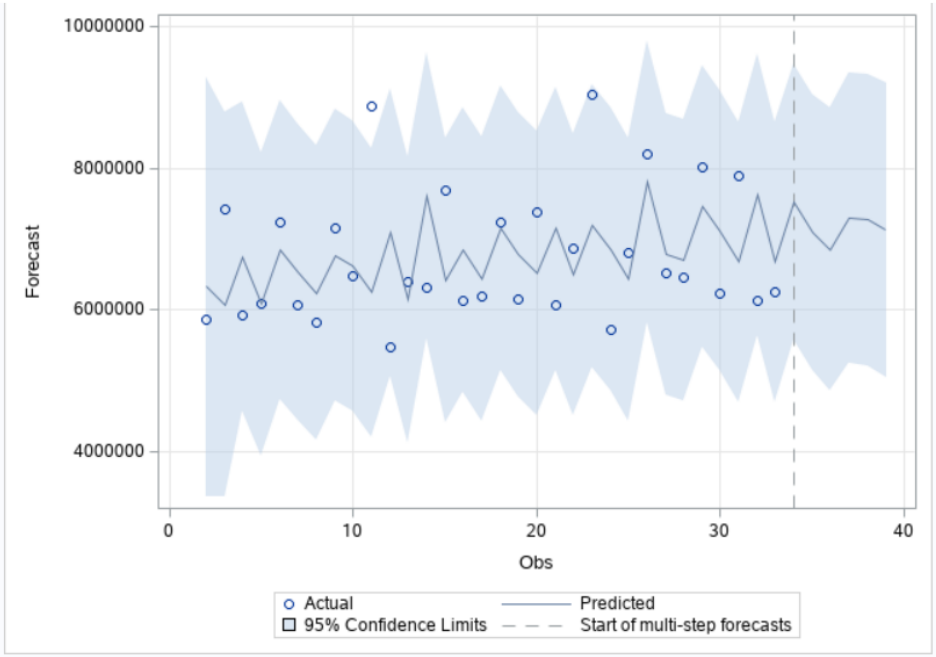


ARIMA Estimation Optimization Summary	
Estimation Method	Maximum Likelihood
Parameters Estimated	5
Termination Criteria	Maximum Relative Change in Estimates
Iteration Stopping Value	0.001
Criteria Value	82.45329
Maximum Absolute Value of Gradient	1.457E13
R-Square Change from Last Iteration	0.620811
Objective Function	Log Gaussian Likelihood
Objective Function Value	-486.43
Marquardt's Lambda Coefficient	0.00001
Numerical Derivative Perturbation Delta	0.001
Iterations	4
Warning Message	Estimates may not have converged.

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	21940.2	19753.7	1.11	0.2667	0
MA1,1	0.99935	67.59411	0.01	0.9882	1
AR1,1	0.04352	0.50762	0.09	0.9317	1
AR1,2	-0.21758	0.38577	-0.56	0.5727	2
AR1,3	0.27069	0.37254	0.73	0.4675	3

Constant Estimate	19819.9
Variance Estimate	9.895E11
Std Error Estimate	994720.7
AIC	982.859
SBC	990.1877
Number of Residuals	32

Correlations of Parameter Estimates					
Parameter	MU	MA1,1	AR1,1	AR1,2	AR1,3
MU	1.000	0.378	0.305	0.211	0.236



Outlier Detection Summary	
Maximum number searched	1
Number found	1
Significance used	0.05

Outlier Details				
Obs	Type	Estimate	Chi-Square	Approx Prob>ChiSq
11	Additive	2773827.1	12.48	0.0004