Sales Forecasting Using SAS For Global Retail Inc.

FINAL PROJECT REPORT

DSCI 5340 – PREDICTIVE ANALYTICS AND BUSINESS FORECASTING

Group Members:
Ashish Kulkarni – ashishkulkarni@my.unt.edu
Marcelo Barbosa Rosalen - marcelobarbosarosalen@my.unt.edu

Submitted to:
Dr. Mahdi Fathi

Table of Contents

Executive Summary

Global Retail (GR) has been enjoying significant challenges in revenue and profit during the last four years, from 2015 to 2018. GR is a retail company spread across worldwide with a presence in all continents. They have three line of business – Home Office, Consumer and Corporate; and three different Category of products – Technology, Office Supplies and Furniture. During the last fiscal year – 2018, each region has a sales manager and are distributed across 24 different regions. The region manager's report sales during the financial year based on Orders and Returns for that specific year.

GR is the largest retail company in the world, with annual revenue of 3.1 billion US dollars during the last four years. The worldwide sales director need to predict a pipeline that provides a monthly forecast for sales numbers.

The GR sales forecasting project is aimed at taking advantage of the information available inside of the company to predict the sales for the next year. The benefit of the forecasting project is evident, especially in the following aspects:

- Accurate and timely forecast in retail business drive success. It's an essential enabler of supply and inventory planning, product pricing, promotion and placement.
- Predictability means profit. The ability to predict will give the company an advantage in all the regions to improve the profit.
- Predictability will enable the company to control and reduce the cost associated with each sale.

Forecasting is a new field that combines business insights with computer learning capabilities. We will use both experiences to explore the models as describe in the next session and find the best strategy to help the GR stakeholders to improve the profit across and the regions.

References:

- Emerald Group Publishing - Forecasting Sales - The foundation of business planning
- Aberdeen Group - How performers leverage the past, visualize the present and improve their future revenue
- Bluepiit - Importance Of Sales Forecasting For Businesses
- Practical Guide to Business Forecasting - Graceway Publishing Company
- Sales forecasting using different forecasting methods by Notkun spálíkana við gerð söluspár

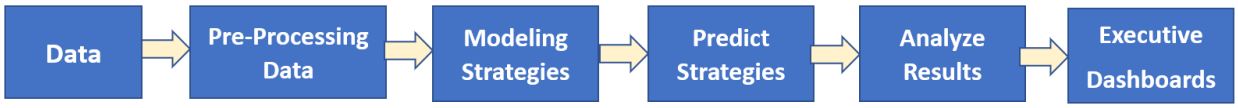
Project Background

As described in the Executive Summary, GR Retail Inc is a company with stores across the globe and presence in all continents. The retail company has three different line of business - Home Office, Consumer and Corporate. **The worldwide sales director needs to build a pipeline that automatically provides sales forecasts for the next year.**

As part of the project strategies, we will use the following models to understand the relationship between a dependent variable **'Sales Amount'** by **'Month'** and understanding how those variables are related.

- Use Excel for data clean up
- Extract relevant data items for forecasting
- Forecasting using SAS
- Exploring the ARIMA Model
- Sales Forecasting using MS Excel

As a stretch exploration to help GR Retail Inc., we will apply the forecasting techniques using the SAS. Using the available methods, we will use available data set to train the model and forecast using the described workflow below.



Data Description and Preparation

The dataset is a second-hand data obtained from Microsoft dataset templates available at Microsoft AI Gallery. The dataset is divided into three different categories, which are:

1. Orders = includes all the data related to sales and orders
2. Returns = includes all the data related to returned orders
3. Persona = includes all the data related to sales manager by region

For this project, we are discarding the 'Returns' and 'Persona' data; and only considering the 'Orders' data. The **Orders** represent the sales by region from the fiscal year 2015 until 2018. The dataset has 51,291 rows and 24 columns as described below:

Data Type	Data Field Name
Numeric	Row ID, Postal Code, Sales, Quantity, Discount, Profit, Shipping Cost,
String	Order ID, Ship Mode, Customer ID, Customer Name, Segment, City, Country, Region, Market, Product ID, Product Name, Sub-Category, Category, Order Priority
Date/Time	Order Date, Ship Date

The dataset Global SuperStore 2018 is related to a Retail company distributed across the globe, and the challenge we are facing is a common business problem of improving business revenue and profit with reduced cost. The objective essentially is to maximize the profit by region while minimizing the cost associated by product segment.

The primary task while initiating an analytics exploration is to define the end goals/objectives of the exploration. A general and straightforward Sales dashboard performed on the dataset show us the numbers presented in Figure 1 – for this purpose, we used a free version of **Microsoft PowerBi**.

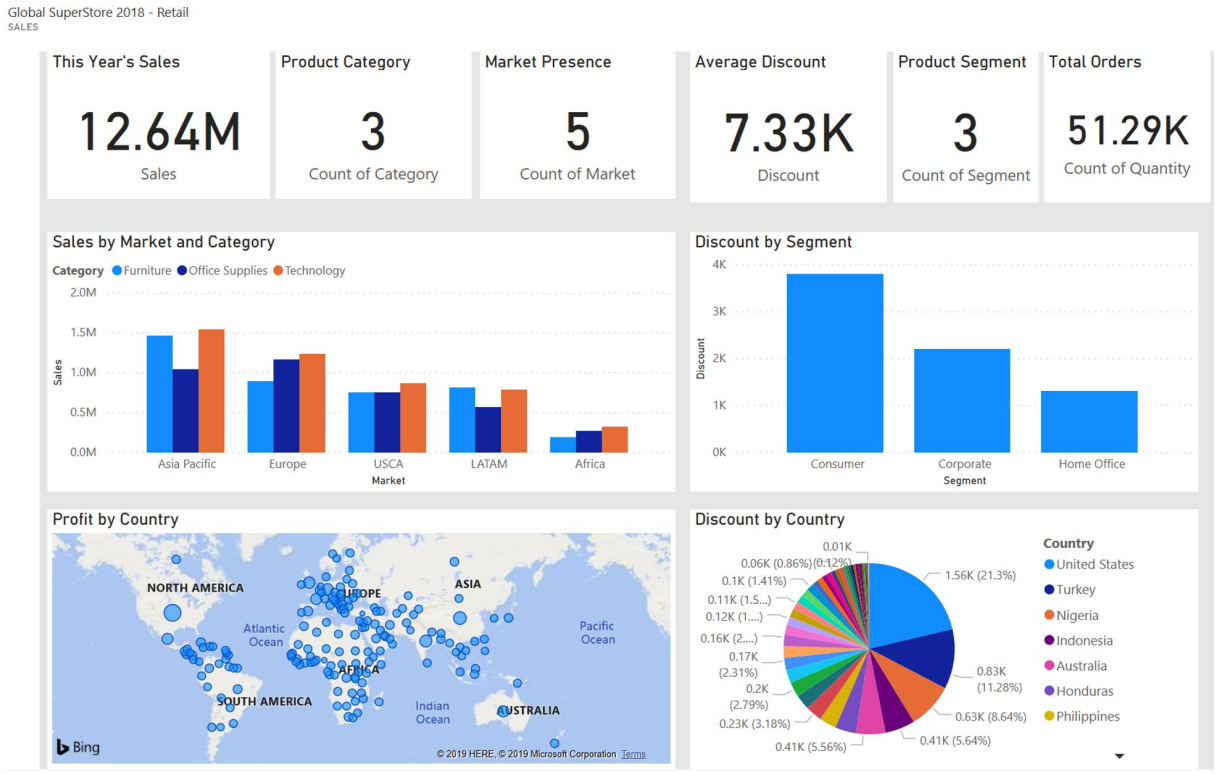


Figure 1 – Sales Dashboard

A quick insight analysis on the Sales metrics shows us that the Sales revenue for the year 2018 was $12.64M across three different Product Category – Furniture, Office Supplies, and Technology; three different Market Segments are – Customer, Corporate and Home Office.

In terms of discount, we can see on the same dashboard that the discount into the Consumer segment is the higher follow by Corporate and Home-Office. But most important and worrisome observation is that the large markets like the USA have the highest discount rate - 21.3%.

Using the same tool and a little quick insight we can see, but not understand why yet, the Profit by Category during the last four years - see Figure 2, and the Profit by Market Segment – see Figure 3

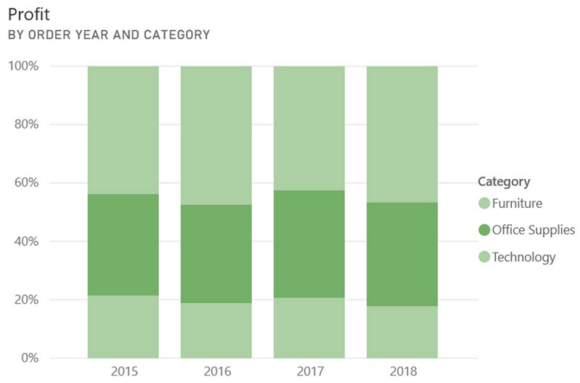


Figure 2 - Profit by Category by Year

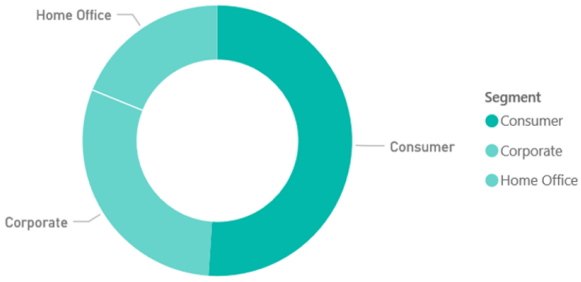
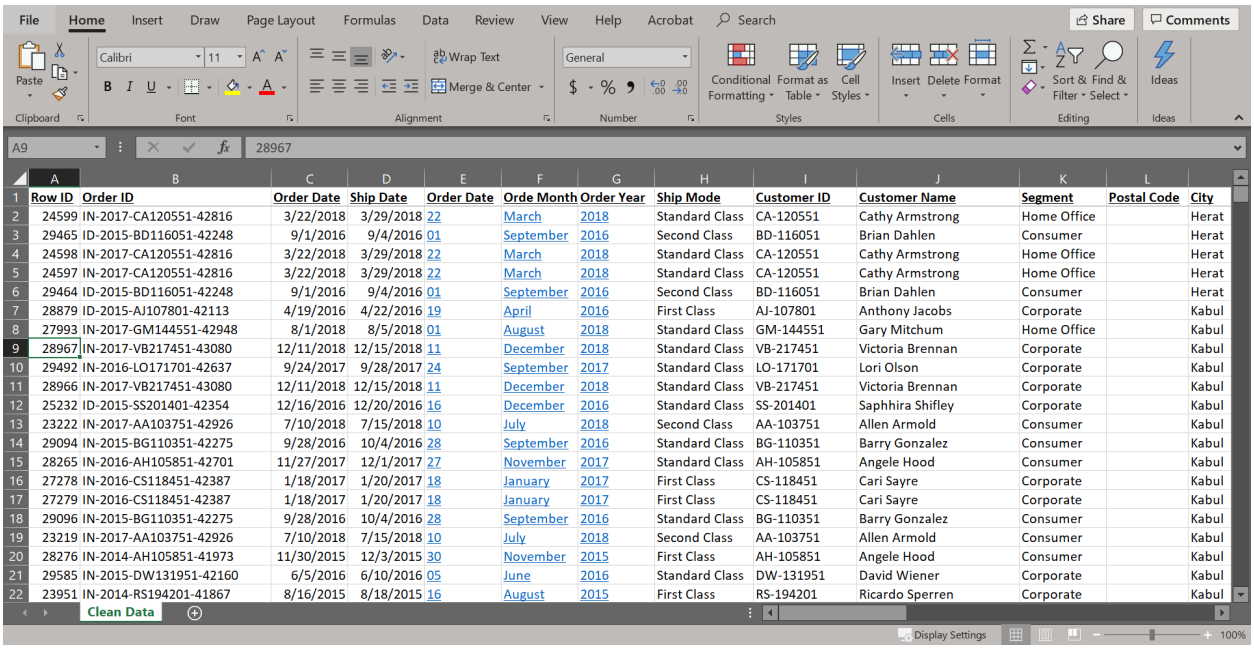


Figure 3 - Profit by Market Segment

Data Clean-Up



Row ID	Order ID	Order Date	Ship Date	Order Date	Orde Month	Order Year	Ship Mode	Customer ID	Customer Name	Segment	Postal Code	City
24599	IN-2017-CA120551-42816	3/22/2018	3/29/2018	22	March	2018	Standard Class	CA-120551	Cathy Armstrong	Home Office		Herat
29465	ID-2015-BD116051-42248	9/1/2016	9/4/2016	01	September	2016	Second Class	BD-116051	Brian Dahlen	Consumer		Herat
24598	IN-2017-CA120551-42816	3/22/2018	3/29/2018	22	March	2018	Standard Class	CA-120551	Cathy Armstrong	Home Office		Herat
24597	IN-2017-CA120551-42816	3/22/2018	3/29/2018	22	March	2018	Standard Class	CA-120551	Cathy Armstrong	Home Office		Herat
29464	ID-2015-BD116051-42248	9/1/2016	9/4/2016	01	September	2016	Second Class	BD-116051	Brian Dahlen	Consumer		Herat
28879	ID-2015-AJ107801-42113	4/19/2016	4/22/2016	19	April	2016	First Class	AJ-107801	Anthony Jacobs	Corporate		Kabul
27993	IN-2017-GM144551-42948	8/1/2018	8/5/2018	01	August	2018	Standard Class	GM-144551	Gary Mitchum	Home Office		Kabul
28967	IN-2017-VB217451-43080	12/11/2018	12/15/2018	11	December	2018	Standard Class	VB-217451	Victoria Brennan	Corporate		Kabul
29492	IN-2016-LO171701-42637	9/24/2017	9/28/2017	24	September	2017	Standard Class	LO-171701	Lori Olson	Corporate		Kabul
28966	IN-2017-VB217451-43080	12/11/2018	12/15/2018	11	December	2018	Standard Class	VB-217451	Victoria Brennan	Corporate		Kabul
25232	ID-2015-SS201401-42354	12/16/2016	12/20/2016	16	December	2016	Standard Class	SS-201401	Saphhira Shifley	Corporate		Kabul
23222	IN-2017-AA103751-42926	7/10/2018	7/15/2018	10	July	2018	Second Class	AA-103751	Allen Armold	Consumer		Kabul
29094	IN-2015-BG110351-42275	9/28/2016	10/4/2016	28	September	2016	Standard Class	BG-110351	Barry Gonzalez	Consumer		Kabul
28265	IN-2016-AH105851-42701	11/27/2017	12/1/2017	27	November	2017	Standard Class	AH-105851	Angele Hood	Consumer		Kabul
27278	IN-2016-CS118451-42387	1/18/2017	1/20/2017	18	January	2017	First Class	CS-118451	Cari Sayre	Corporate		Kabul
27279	IN-2016-CS118451-42387	1/18/2017	1/20/2017	18	January	2017	First Class	CS-118451	Cari Sayre	Corporate		Kabul
29096	IN-2015-BG110351-42275	9/28/2016	10/4/2016	28	September	2016	Standard Class	BG-110351	Barry Gonzalez	Consumer		Kabul
23219	IN-2017-AA103751-42926	7/10/2018	7/15/2018	10	July	2018	Second Class	AA-103751	Allen Armold	Consumer		Kabul
28276	IN-2014-AH105851-41973	11/30/2015	12/3/2015	30	November	2015	First Class	AH-105851	Angele Hood	Consumer		Kabul
29585	IN-2015-DW131951-42160	6/5/2016	6/10/2016	05	June	2016	Standard Class	DW-131951	David Wiener	Corporate		Kabul
23951	IN-2014-RS194201-41867	8/16/2015	8/18/2015	16	August	2015	First Class	RS-194201	Ricardo Sperren	Corporate		Kabul

During data exploration in Python, it was found that the data set had an empty column named Postal_Code. This empty column was then eliminated.

```
#validation of data
datanull = dataset[dataset.isnull().any(axis=1)]
print(dataset.isnull().sum())
datanull.head()
```

```
Postal Code      41296
City                 0
State                0
```

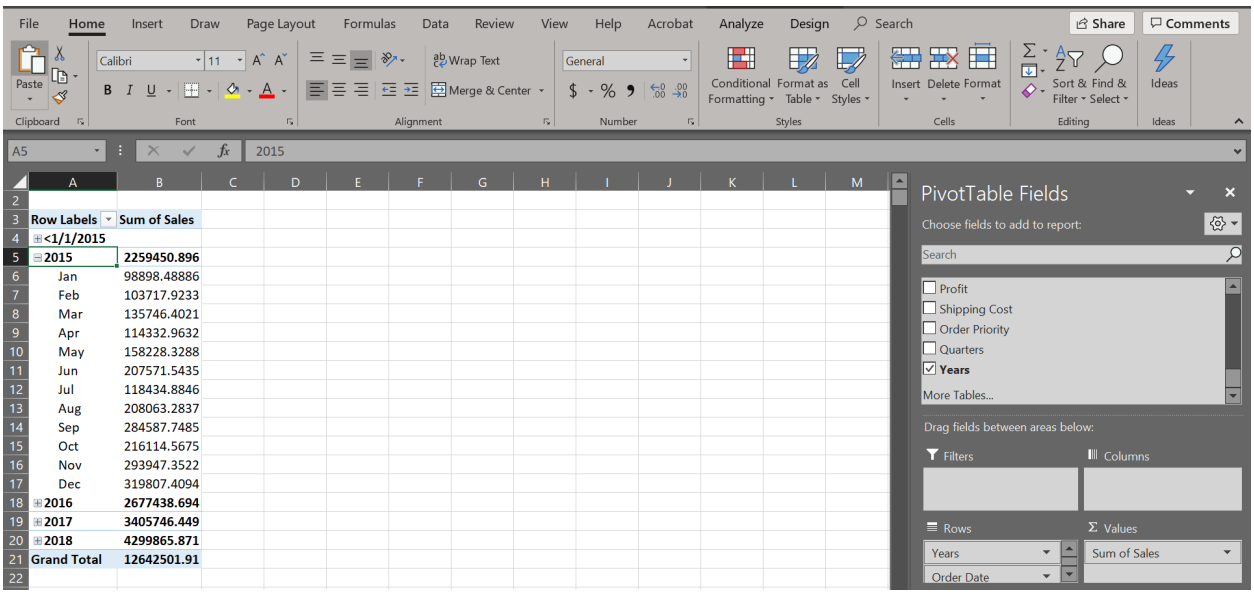
```
#Data has no missing values except for postal codes.
#Postal code has no relevance in our data analysis, hence deleting postal code column
cols = ['Postal Code']
dataset.drop(cols, axis=1, inplace=True)
```

It is now confirmed that the dataset has no missing values.

Extract Relevant Data Items for Forecasting

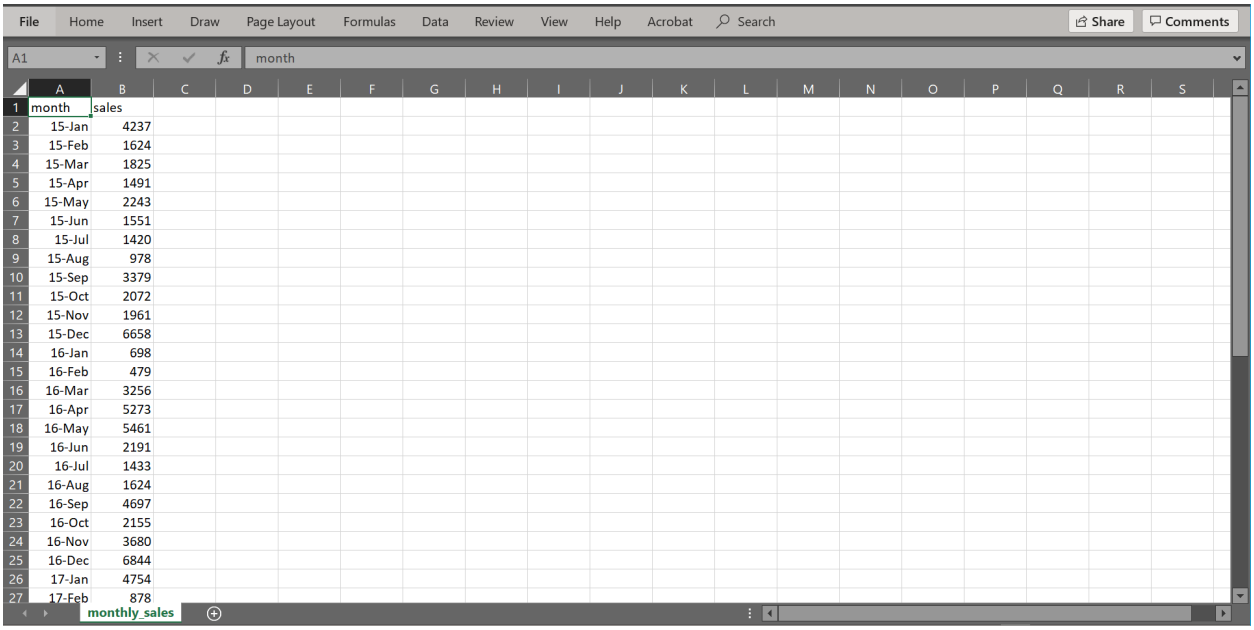
Dataset has 27 columns and 51,291 rows. Considering the expected output, monthly sales forecasting not all these columns are useful during the analysis. Also, the data is available at the day level of granularity. It was decided to ignore unnecessary columns and use monthly sales data instead of daily.

Daily sales data was converted to monthly sales data. This data was used for further forecasting.



Monthly Sales Data:

A new dataset in CSV format was created that had only two data columns. Month and Sales. This time-series data will help to forecast the sales values for an additional year in advance.



month	sales
15-Jan	4237
15-Feb	1624
15-Mar	1825
15-Apr	1491
15-May	2243
15-Jun	1551
15-Jul	1420
15-Aug	978
15-Sep	3379
15-Oct	2072
15-Nov	1961
15-Dec	6658
16-Jan	698
16-Feb	479
16-Mar	3256
16-Apr	5273
16-May	5461
16-Jun	2191
16-Jul	1433
16-Aug	1624
16-Sep	4697
16-Oct	2155
16-Nov	3680
16-Dec	6844
17-Jan	4754
17-Feb	878

Forecasting

The FORECAST procedure provides a quick and automatic way to generate forecasts for many time series in one step. The procedure can forecast hundreds of series at a time, with the series organized into separate variables or across BY groups. PROC FORECAST uses extrapolative forecasting methods where the forecasts for a series are functions only of time and past values of the series, not of other variables. [Ref: SAS/ETS® 13.2 User's Guide The FORECAST Procedure].

The FORECAST procedure writes the forecasts and confidence limits to an output data set. It can also write parameter estimates and fit statistics to an output data set. This dataset will be printed to understand the estimated forecast values.

Method

Three different forecasting methods were used, and their outputs were compared.

SAS Forecasting

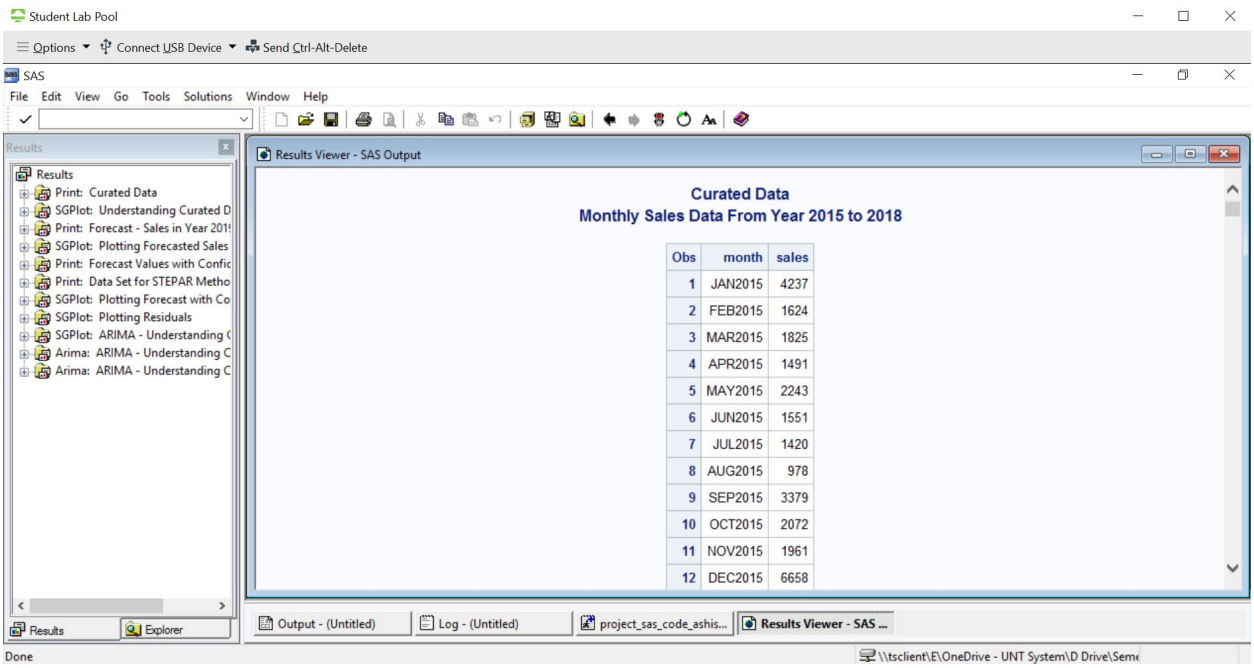
The below code was written and used in SAS to perform the forecasting.

```
/*Importing dataset - monthly sales numbers from year 2015 to 2018*/
PROC IMPORT OUT=WORK.SalesData
DATAFILE= "C:\temp\monthly_sales.csv"
DBMS=csv REPLACE;
GETNAMES=Yes;
RUN;

/*Printing the imported dataset*/
title1 'Curated Data';
title2 'Monthly Sales Data From Year 2015 to 2018';
proc print data=SalesData;
run;

title1 'Understanding Curated Data';
proc sgplot data=SalesData;
scatter y=sales x=month;
run;

/*Forecast for next year, i.e. 2019*/
title1 'Forecast - Sales in Year 2019';
title2 'Monthly Forecasted Sales Data For Year 2019';
proc forecast data=SalesData interval=month lead=12 out=prediction;
var sales;
id month;
run;

/*printing forecast values for year 2019*/
proc print data=prediction;
run;

/*Forecast with confidence limit*/
title1 'Plotting Forecasted Sales For Year 2019';
title2 'Based on Sales Data From Year 2015 to 2018';
proc sgplot data=prediction;
series x=month y=sales / markers markerattrs=(symbol=circlefilled);
xaxis values=('1jan19'd to '1dec19'd by mon);
label month = 'Months in Year 2019';
label sales = 'Forecasted Sales in Year 2019';
run;

/*printing forecast values with confidence limit*/
```

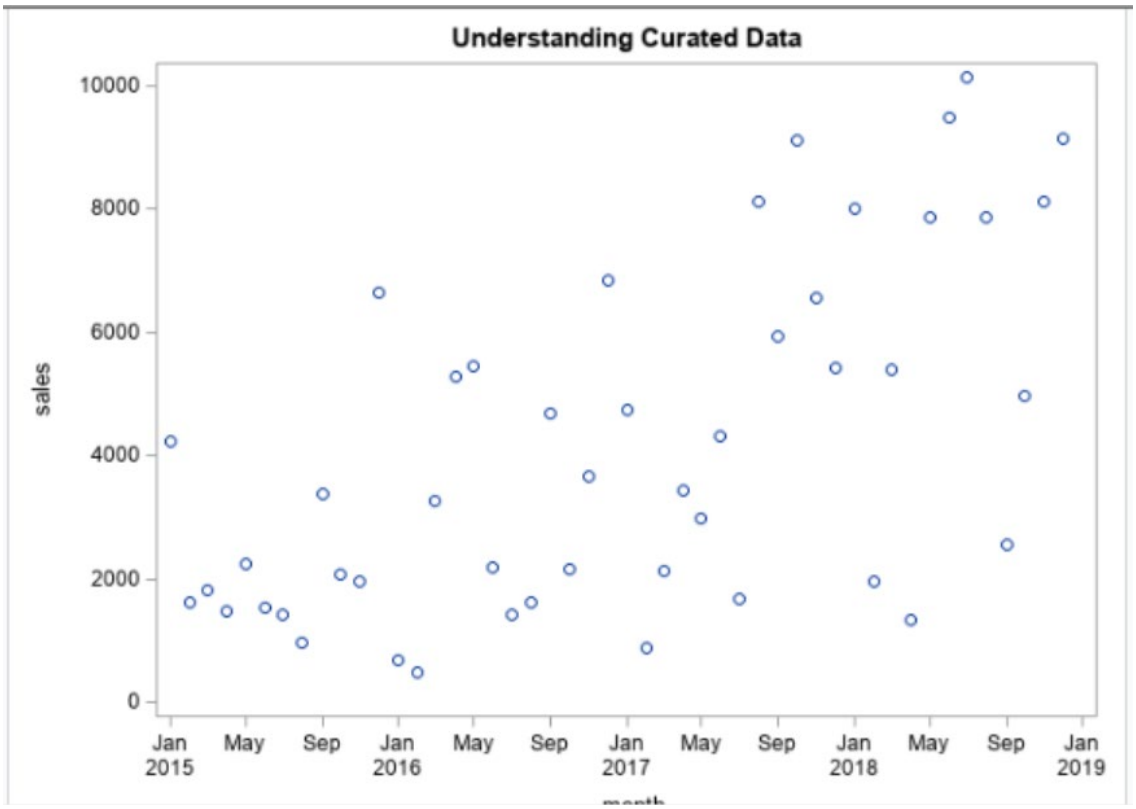
```
title1 'Forecast Values with Confidence Limit';
proc forecast data=SalesData interval=month lead=12
       out=pred
       OUTFULL
       OUTRESID
       OUTEST=EST
       OUTFITSTATS;
var sales;
id month;
run;

proc print data=pred;
run;

/*Model Parameters and Goodness-of-Fit Statistics*/
title1 'Data Set for STEPAR Method';
proc print data=est;
run;

/*Plotting forecast*/
title1 'Plotting Forecast with Confidence Limits';
proc sgplot data=pred;
series x=month y=sales / group=_type_ lineattrs=(pattern=1);
xaxis values=('1jan15'd to '1dec19'd by mon);
refline '1jan2019'd / axis=x;
run;

/*Plotting Residuals*/
title1 'Plotting Residuals';
proc sgplot data=pred;
where _type_='RESIDUAL';
needle x=month y=sales / markers;
xaxis values=('1jan15'd to '1dec18'd by mon);
run;
```

Output

Printing and understanding imported data.



Obs	month	sales
1	JAN2015	4237
2	FEB2015	1624
3	MAR2015	1825
4	APR2015	1491
5	MAY2015	2243
6	JUN2015	1551
7	JUL2015	1420
8	AUG2015	978
9	SEP2015	3379
10	OCT2015	2072
11	NOV2015	1961
12	DEC2015	6658



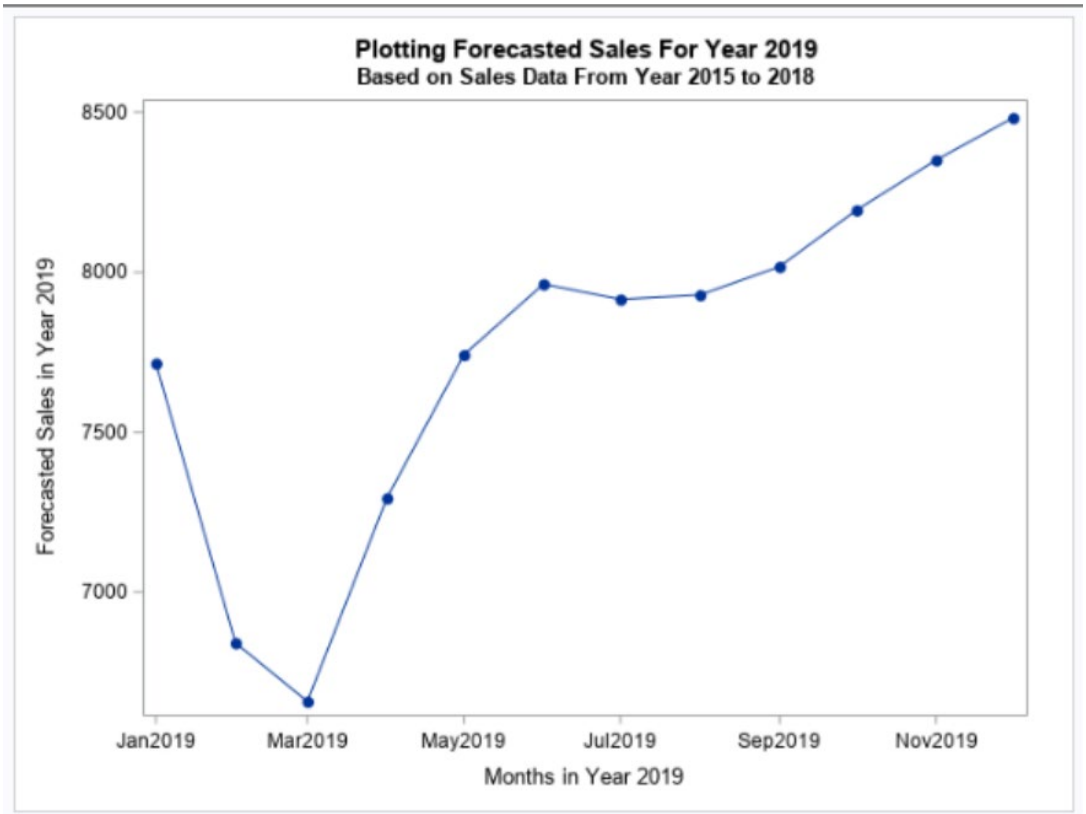
The below table illustrates the forecasted sales numbers for the next 12 months. The ingested data is from the year 2015 to 2018, the forecasting is done for the year 2019, and month-wise data can be seen below.

Forecast - Sales in Year 2019
Monthly Forecasted Sales Data For Year 2019

Obs	month	_TYPE_	_LEAD_	sales
1	JAN2019	FORECAST	1	7715.3578531
2	FEB2019	FORECAST	2	6841.3649163
3	MAR2019	FORECAST	3	6658.2371929
4	APR2019	FORECAST	4	7293.9167955
5	MAY2019	FORECAST	5	7741.8908118
6	JUN2019	FORECAST	6	7963.0137145
7	JUL2019	FORECAST	7	7915.2746884
8	AUG2019	FORECAST	8	7929.1702917
9	SEP2019	FORECAST	9	8017.5542708
10	OCT2019	FORECAST	10	8194.2212128
11	NOV2019	FORECAST	11	8350.6499318
12	DEC2019	FORECAST	12	8482.6197972

Plotting the Forecast

The provided forecast was plotted using the **sgplot** method. One can see that the predicted sales fall in the first two years of the new year and then bounce backs. These forecast numbers will help warehouse, logistics, store managers, and last-mile delivery agents to plan their resources appropriately.



Forecast Values with Confidence Limit

Values were forecasted with an upper and lower level at 95% confidence for all 12 months. For this, the below options were specified:

- OUTFULL
- OUTRESID
- OUTEST=EST
- OUTFITSTATS

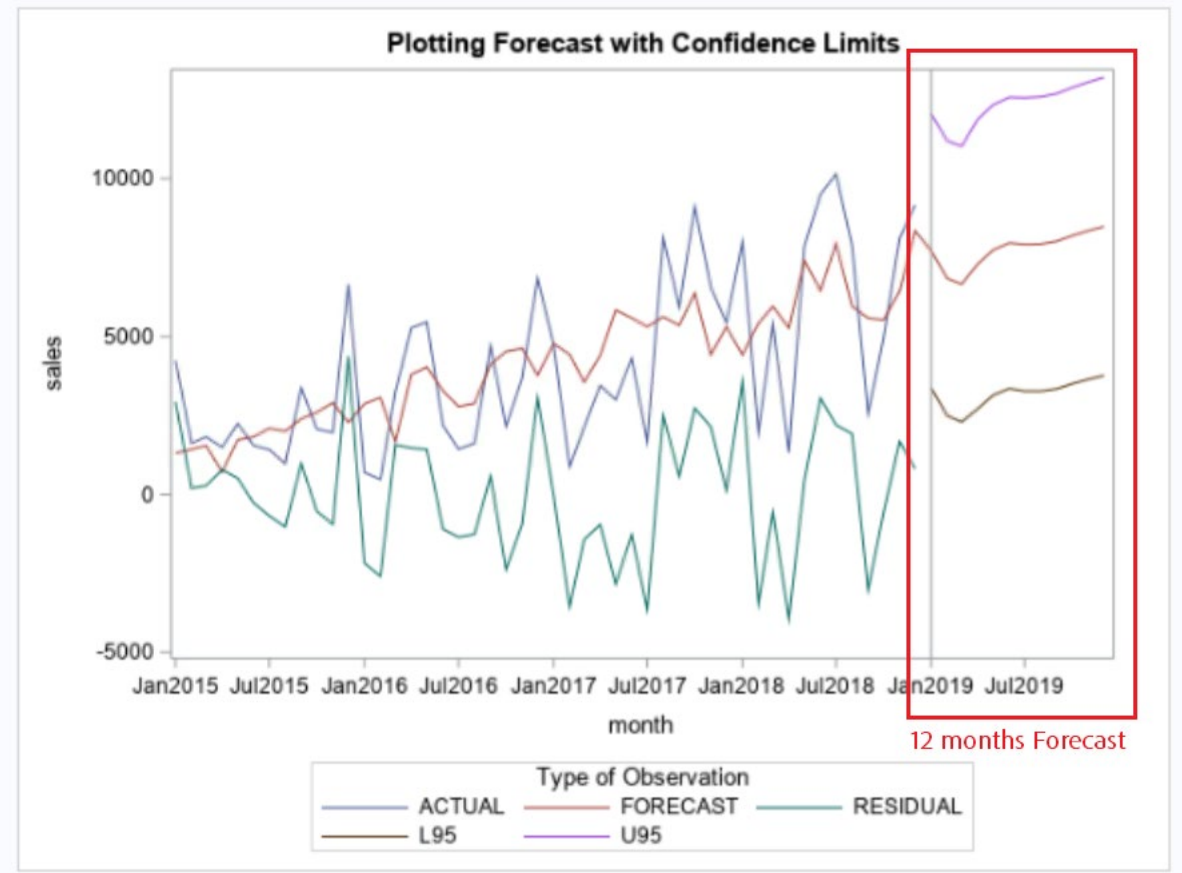
Forecast Values with Confidence Limit

Obs	month	_TYPE_	_LEAD_	sales
1	JAN2015	ACTUAL	0	4237
2	JAN2015	FORECAST	0	1306.9812925
3	JAN2015	RESIDUAL	0	2930.0187075
4	FEB2015	ACTUAL	0	1624
5	FEB2015	FORECAST	0	1428.177124
6	FEB2015	RESIDUAL	0	195.82287596
7	MAR2015	ACTUAL	0	1825
8	MAR2015	FORECAST	0	1549.3729556
9	MAR2015	RESIDUAL	0	275.62704443
10	APR2015	ACTUAL	0	1491
11	APR2015	FORECAST	0	708.47373027
12	APR2015	RESIDUAL	0	782.52626973
13	MAY2015	ACTUAL	0	2243
14	MAY2015	FORECAST	0	1727.4646124
15	MAY2015	RESIDUAL	0	515.53538756

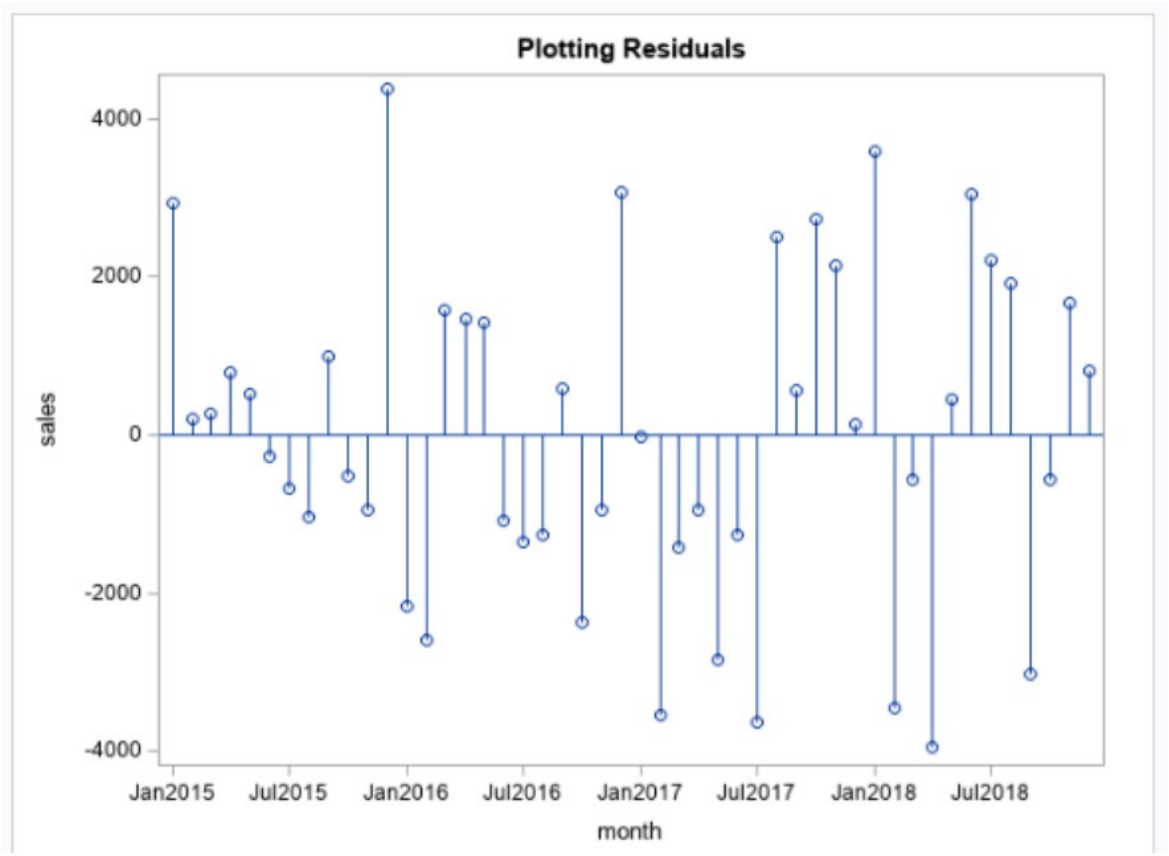
Data Set for STEPAR Method

Obs	_TYPE_	month	sales
1	N	DEC2018	48
2	NRESID	DEC2018	48
3	DF	DEC2018	45
4	SIGMA	DEC2018	2114.2593
5	CONSTANT	DEC2018	1185.7855
6	LINEAR	DEC2018	121.19583
7	AR01	DEC2018	.
8	AR02	DEC2018	.
9	AR03	DEC2018	-0.328358
10	AR04	DEC2018	.
11	AR05	DEC2018	.
12	AR06	DEC2018	.
13	AR07	DEC2018	.
14	AR08	DEC2018	.

Plotting Forecast with Confidence Limits:



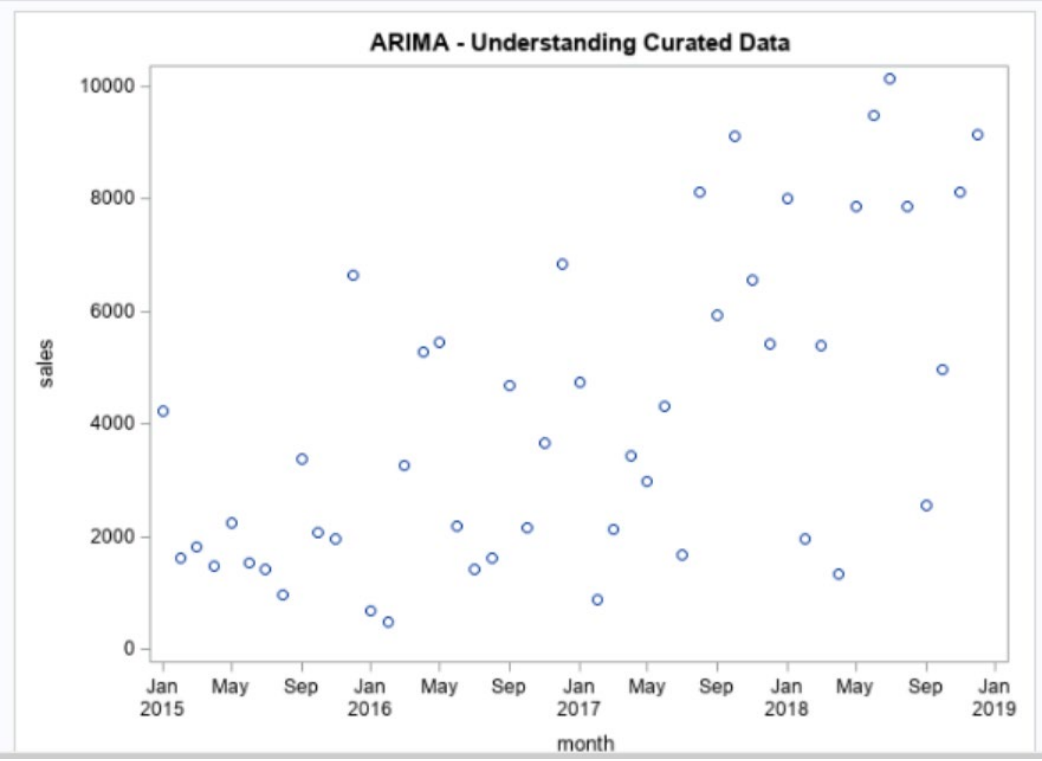
Plotting Residuals:


Plotting Residuals
4000
2000
0
-2000
-4000
sales
Jan2015
Jul2015
Jan2016
Jul2016
Jan2017
Jul2017
Jan2018
Jul2018
month

ARIMA Method of Forecasting

The ARIMA procedure analyzes and forecasts equally spaced univariate time series data, transfer function data, and intervention data by using the autoregressive integrated moving-average (ARIMA) or autoregressive moving-average (ARMA) model. An ARIMA model predicts a value in a response time series as a linear combination of its past values, past errors (also called shocks or innovations), and current and past values of other time series.

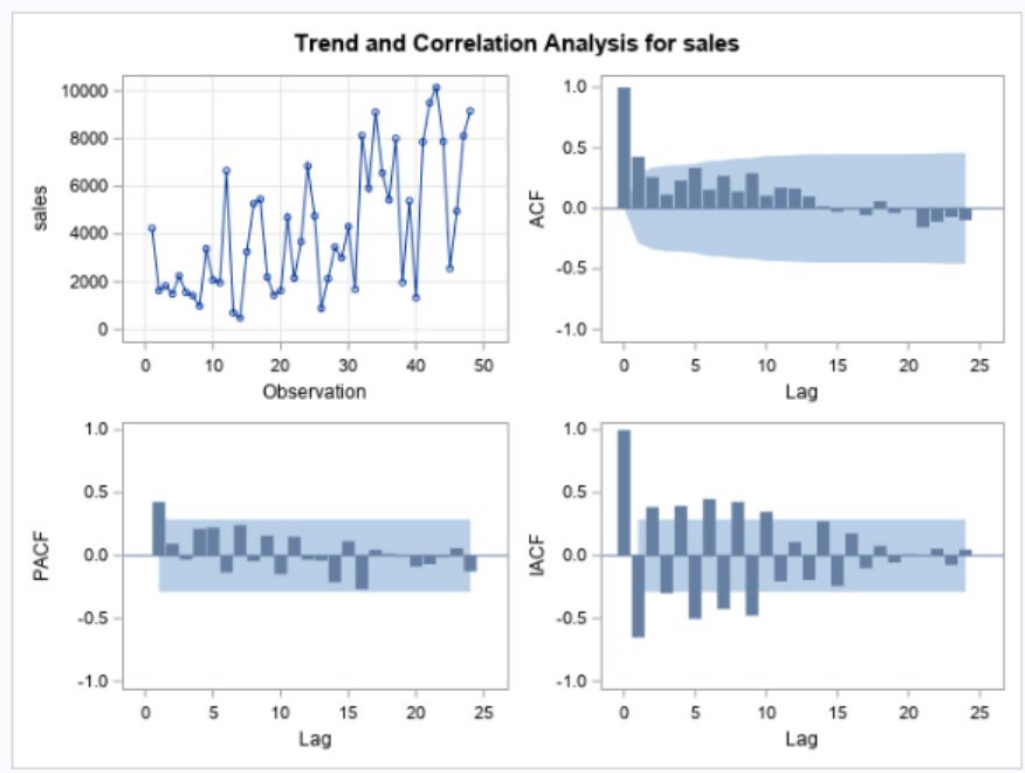
Box and Jenkins first popularized the ARIMA approach, and ARIMA models are often referred to as Box-Jenkins models. The general transfer function model employed by the ARIMA procedure was discussed by Box and Tiao (1975). When an ARIMA model includes other time series as input variables, the model is sometimes referred to as an ARIMAX model. Pankratz (1991) refers to the ARIMAX model as dynamic regression. [Ref: SAS/ETS® 13.2 User's Guide The ARIMA Procedure]



Name of Variable = sales	
Mean of Working Series	4155.083
Standard Deviation	2741.552
Number of Observations	48

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	24.10	6	0.0005	0.428	0.260	0.115	0.230	0.335	0.156
12	39.26	12	<.0001	0.270	0.141	0.292	0.105	0.174	0.165
18	40.53	18	0.0018	0.099	0.017	-0.027	-0.007	-0.055	0.060
24	45.35	24	0.0053	-0.038	0.004	-0.157	-0.110	-0.071	-0.096

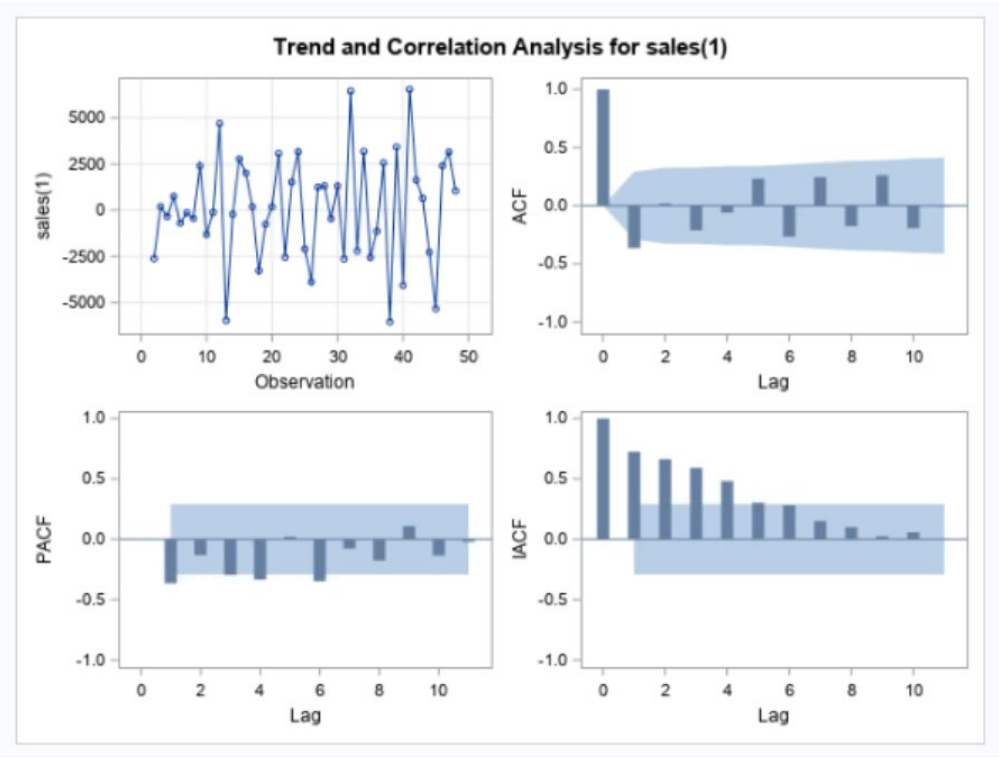
Trend and Correlation Analysis for Sales:



Name of Variable = sales	
Period(s) of Differencing	1
Mean of Working Series	104.766
Standard Deviation	2871.085
Number of Observations	47
Observation(s) eliminated by differencing	1

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	16.17	6	0.0129	-0.365	0.019	-0.212	-0.060	0.232	-0.267

Trend and Correlation Analysis:

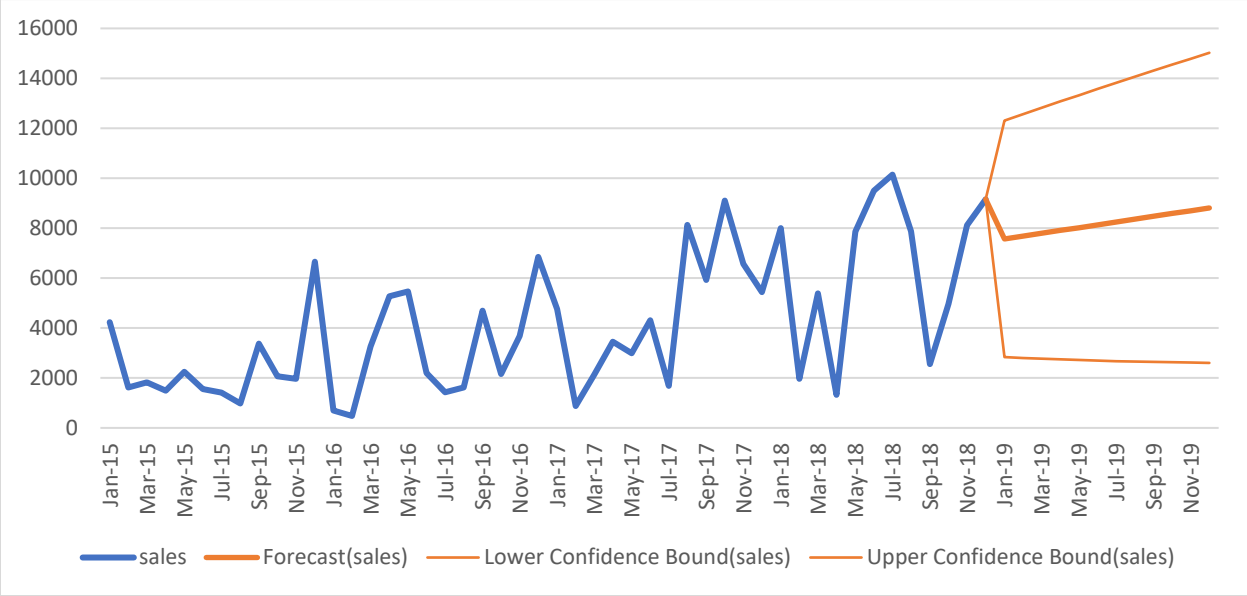


Forecasting for Sales:

Forecasts for variable sales				
Obs	Forecast	Std Error	95% Confidence Limits	
49	7553.1399	2447.6779	2755.7794	12350.5005
50	7270.9016	2572.0822	2229.7132	12312.0899
51	7276.2731	2607.8835	2164.9155	12387.6308
52	7344.0452	2631.6147	2186.1751	12501.9153
53	7425.3558	2652.9389	2225.6912	12625.0205
54	7509.6038	2673.6344	2269.3767	12749.8310
55	7594.4891	2694.0729	2314.2033	12874.7750
56	7679.5127	2714.3364	2359.5111	12999.5143
57	7764.5662	2734.4452	2405.1521	13123.9804
58	7849.6263	2754.4063	2451.0892	13248.1634
59	7934.6878	2774.2235	2497.3097	13372.0658
60	8019.7496	2793.9001	2543.8061	13495.6930

Using MS-Excel for forecasting

If you have historical time-based data, you can use MS Excel to create a forecast. When you create a forecast, Excel creates a new worksheet that contains both a table of the historical and predicted values and a chart that expresses this data.

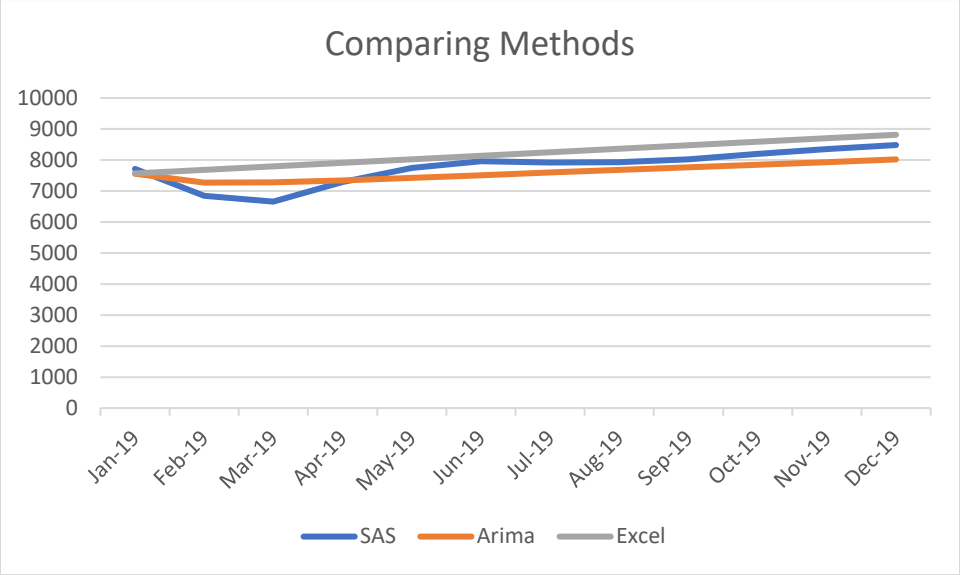


month	sales	Forecast(sales)	Lower Confidence Bound(sales)	Upper Confidence Bound(sales)
Nov-18	8113			
Dec-18	9161	9161	9161	9161
Jan-19		7568	2833	12302
Feb-19		7681	2800	12562
Mar-19		7794	2769	12819
Apr-19		7907	2742	13073
May-19		8021	2717	13324
Jun-19		8134	2694	13573
Jul-19		8247	2674	13820
Aug-19		8360	2656	14064
Sep-19		8473	2640	14307
Oct-19		8587	2626	14547
Nov-19		8700	2613	14786
Dec-19		8813	2602	15024

Comparing Different Methods:

When we compared the forecast values of three different models, we found that the though results were not same (which was also not an expectation) they fall in-line, and sales were predicted to increase in all the models.

Hence, it is safe to assume that the sale will move in an upward direction.


Comparing Methods
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Jan-19
Feb-19
Mar-19
Apr-19
May-19
Jun-19
Jul-19
Aug-19
Sep-19
Oct-19
Nov-19
Dec-19
SAS
Arima
Excel

Conclusion

The project involved working with a dataset provided by a retail organization and group was tasked to provide a sales forecast for the next 12 months considering sales data for the previous 48 months. SAS was used to forecast the sales numbers.

From the forecasting exercise, it is concluded that the sales numbers will go up in the next 12 months, and a month-wise forecast was provided to the organization. Using this data organization can ensure that their manufacturing, procurement, sales, marketing, and distribution teams are prepared well in advance for deeps and spikes in sales numbers.

Having actionable insights about the sales numbers can prove to be of immense value, and here, our group aimed to provide the same.

What Did We Learn?

As a final comment on this project, as a group, we can say with confidence that we have demonstrated a vast learning experience across all the phases of the project and methods used on it. As a future data scientist, it was an awesome experience to go over each step, including the process to understand the business problem, searching for a useful dataset.

Starting on the business problem step and the definition of data mining strategies, we had an excellent insight to use a Global Retail store data set that help us during the stages to apply and compare the methods. At the same time, as a result of this step, it was easy to understand the business problem and identify which business question we would like to resolve.

On the forecasting step, we decided to use three different methods, that could have been a risk to the group, but at the end was a fantastic learning experience beyond the classroom.

In fact, above all the steps together helped us to understand how a competent data mining strategy can help the organization of every size which has a clear vision and goals in mind based on the dataset available to them.

== END OF REPORT ==