SUPERSTORE SALES DATASET ANALYSIS AND FORECASTING

A Time Series Project Report

Gourab Ghosh
M.Sc. in Big Data Analytics, Department of Computer Science
Ramakrishna Mission Vivekananda Educational And Research Institute
Belur Math, Howrah
Pin-711202, West Bengal
January 03, 2021

ACKNOWLEDGEMENT

THIS IS TO ACKNOWLEDGE THE SUPPORT AND HELP THAT HAS BEEN RECEIVED WHILE DOING THIS PROJECT, **SUPERSTORE SALES DATASET ANALYSIS AND FORECASTING** FROM OUR TEACHER, **Dr. Sudipta Das**. IT IS VERY KIND OF HIM TO HELP ME IN VARIOUS STAGES WHILE DOING THE PROJECT AND TO PROVIED NECESSARY INFORMATIONS WHICH ARE OF GEART HELP. I AM VERY MUCH GRATEFUL TO HIM AND THANKS HIM FOR HIS CORDIAL COORDINATION.

I WOULD ALSO LIKE TO THANK MY INSTITUTION AND FACULTY MEMBERS FOR HELPING ME IN MY PROJECT. I WANT TO EXPRESS MY GRATITUDE TO BR.TAMAL MJ, DR. ADITYA BAGCHI AND SWATHY PRABHU MAHARAJ, H.O.D OF COMPUTER SCIENCE DEPT, RKMVERI BELUR, FOR EXTENDING THEIR SUPPORT.

AT LAST BUT NOT THE LEAST I WOULD LIKE TO EXPRESS MY THANKS TO MY PARENTS WITHOUT WHOM THIS PROJECT WOULD NOT HAVE BEEN POSSIBLE.

A **Project Report** SUBMITTED BY **GOURAB GHOSH**
TO RAMAKRISHNA MISSION VIVEKANANDA EDUCATIONAL AND RESEARCH INSTITUTE FOR THE COURSE OF **TIME SERIES AND FORECASTING** IN **M.Sc. in Big Data Analytics**
UNDER THE PROPER GUIDANCE OF **Dr. Sudipta Das**

TABLE OF CONTENT

1 INTRODUCTION

A **time series** is a series of data points indexed or listed or graphed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time.
Time Series analysis can be useful to see how a given asset, security or economic variable changes over time. Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.
A number of different notations are in use for time-series analysis. A common notation specifying a **time series X** that is indexed by the natural numbers is written as **X = [X1, X2, ...]**.Another common notation is **Y = [Yt: t belongs to T]**, where **T is the index set**.

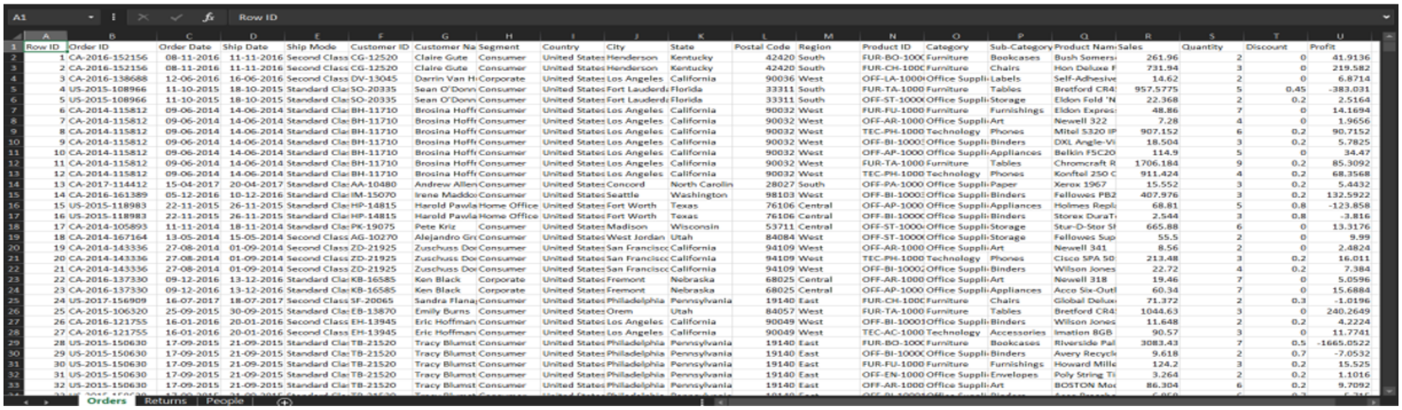
2 OVERVIEW OF PROJECT

Superstore Sales Data is a dataset of a multistore in United States. Here we have different columns, of which our time series analysis interest lies in **category column**, which is of **3 types**, **Furniture, Office Supplies and Technology** with **Sales column**.
For analysis and forecasting of time series, we have to go through following stages:

• At first, we have to read the dataset and check if there is any missing value in it or not and depending on that we have to prepare that accordingly.
• Then **Exploratory Data Analysis** has to be done and basis on that we will try to extract some meaningful information.
• After Eda, we have to do **Data Processing** and prepare the variables, **Furniture, Office Supplies and Technology** for time series analysis accordingly.
• We have to **visualize our dataset** then and try to get meaningful idea and important vision to work on.
• For time series, an important part is **Stationarity Checking**. We have to then check the Stationarity of our time series data and make them stationary if not.
• After Stationarity Checking, **ACF PACF plots** will be drawn and from there we can get the idea of our time series model and their components.
• Then comes the Forecasting part, where we will **forecast our time series** and based on different fitted models and their **AIC values**, we can choose our best time series model.
• We will then visualize our forecasted results and will try to get the forecast for next few time steps.
• At the end, we will try to check if there is any **causal relationship between the time series** and will do **the Granger's Causality Test** for that.

3 DATA COLLECTION

We are going to work on **Superstore Sales Dataset**. An overview of the dataset is like as follows:

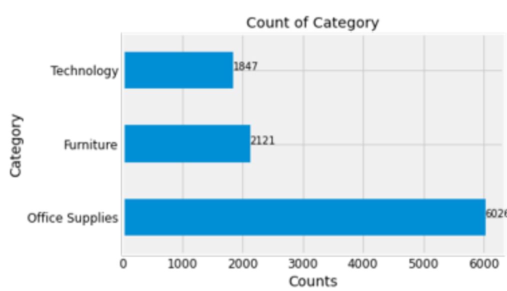


Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Na	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Nam	Sales	Quantity	Discount	Profit
1	CA-2016-152156	08-11-2016	11-11-2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-BO-100C	Furniture	Bookcases	Bush Somers	261.96	2	0	41.9136
2	CA-2016-152156	08-11-2016	11-11-2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-100C	Furniture	Chairs	Hon Deluxe F	731.94	3	0	219.582
3	CA-2016-138688	12-06-2016	16-06-2016	Second Class	DV-13045	Darrin Van H	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-1000	Office Suppli	Labels	Self-Adhesive	14.62	2	0	6.8714
4	US-2015-108966	11-10-2015	18-10-2015	Standard Cla	SO-20335	Sean O'Donn	Consumer	United States	Fort Lauderd	Florida	33311	South	FUR-TA-1000	Furniture	Tables	Bretford CR4	957.5775	5	0.45	-383.031
5	US-2015-108966	11-10-2015	18-10-2015	Standard Cla	SO-20335	Sean O'Donn	Consumer	United States	Fort Lauderd	Florida	33311	South	OFF-ST-1000	Office Suppli	Storage	Eldon Fold 'N	22.368	2	0.2	2.5164
6	CA-2014-115812	09-06-2014	14-06-2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-1000	Furniture	Furnishings	Eldon Expres	48.86	7	0	14.1694
7	CA-2014-115812	09-06-2014	14-06-2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-1000	Office Suppli	Art	Newell 322	7.28	4	0	1.9656
8	CA-2014-115812	09-06-2014	14-06-2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-1000	Technology	Phones	Mitel 5320 IP	907.152	6	0.2	90.7152
9	CA-2014-115812	09-06-2014	14-06-2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-1000	Office Suppli	Binders	DXL Angle-Vi	18.504	3	0.2	5.7825
10	CA-2014-115812	09-06-2014	14-06-2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-1000	Office Suppli	Appliances	Belkin F5C20	114.9	5	0	34.47
11	CA-2014-115812	09-06-2014	14-06-2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-1000	Furniture	Tables	Chromcraft R	1706.184	9	0.2	85.3092
12	CA-2014-115812	09-06-2014	14-06-2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-1000	Technology	Phones	Konftel 250 C	911.424	4	0.2	68.3568
13	CA-2017-114412	15-04-2017	20-04-2017	Standard Cla	AA-10480	Andrew Allen	Consumer	United States	Concord	North Carolin	28027	South	OFF-PA-1000	Office Suppli	Paper	Xerox 1967	15.552	3	0.2	5.4432
14	CA-2016-161389	05-12-2016	10-12-2016	Standard Cla	IM-15070	Irene Maddo	Consumer	United States	Seattle	Washington	98103	West	OFF-BI-1000	Office Suppli	Binders	Fellowes PB2	407.976	3	0.2	132.5922
15	US-2015-118983	22-11-2015	26-11-2015	Standard Cla	HP-14815	Harold Pawla	Home Office	United States	Fort Worth	Texas	76106	Central	OFF-AP-1000	Office Suppli	Appliances	Holmes Repl	68.81	5	0.8	-123.858
16	US-2015-118983	22-11-2015	26-11-2015	Standard Cla	HP-14815	Harold Pawla	Home Office	United States	Fort Worth	Texas	76106	Central	OFF-BI-1000	Office Suppli	Binders	Storex DuraT	2.544	3	0.8	-3.816
17	CA-2014-105893	11-11-2014	18-11-2014	Standard Cla	PK-19075	Pete Kriz	Consumer	United States	Madison	Wisconsin	53711	Central	OFF-ST-1000	Office Suppli	Storage	Stur-D-Stor Sl	665.88	6	0	13.3176
18	CA-2014-167164	13-05-2014	15-05-2014	Second Class	AG-10270	Alejandro Gr	Consumer	United States	West Jordan	Utah	84084	West	OFF-ST-1000	Office Suppli	Storage	Fellowes Sup	55.5	2	0	9.99
19	CA-2014-143336	27-08-2014	01-09-2014	Second Class	ZD-21925	Zuschuss Do	Consumer	United States	San Francisc	California	94109	West	OFF-AR-1000	Office Suppli	Art	Newell 341	8.56	2	0	2.4824
20	CA-2014-143336	27-08-2014	01-09-2014	Second Class	ZD-21925	Zuschuss Do	Consumer	United States	San Francisc	California	94109	West	TEC-PH-1000	Technology	Phones	Cisco SPA 50	213.48	3	0.2	16.011
21	CA-2014-143336	27-08-2014	01-09-2014	Second Class	ZD-21925	Zuschuss Do	Consumer	United States	San Francisc	California	94109	West	OFF-BI-1000	Office Suppli	Binders	Wilson Jones	22.72	4	0.2	7.384
22	CA-2016-137330	09-12-2016	13-12-2016	Standard Cla	KB-16585	Ken Black	Corporate	United States	Fremont	Nebraska	68025	Central	OFF-AR-1000	Office Suppli	Art	Newell 318	19.46	7	0	5.0596
23	CA-2016-137330	09-12-2016	13-12-2016	Standard Cla	KB-16585	Ken Black	Corporate	United States	Fremont	Nebraska	68025	Central	OFF-AP-1000	Office Suppli	Appliances	Acco Six-Outl	60.34	7	0	15.6884
24	US-2017-156909	16-07-2017	18-07-2017	Second Class	SF-20065	Sandra Flana	Consumer	United States	Philadelphia	Pennsylvania	19140	East	FUR-CH-100C	Furniture	Chairs	Global Delux	71.372	2	0.3	-1.0196
25	CA-2015-106320	25-09-2015	30-09-2015	Standard Cla	EB-13870	Emily Burns	Consumer	United States	Orem	Utah	84057	West	FUR-TA-1000	Furniture	Tables	Bretford CR4	1044.63	3	0	240.2649
26	CA-2016-121755	16-01-2016	20-01-2016	Second Class	EH-13945	Eric Hoffman	Consumer	United States	Los Angeles	California	90049	West	OFF-BI-1000	Office Suppli	Binders	Wilson Jones	11.648	2	0.2	4.2224
27	CA-2016-121755	16-01-2016	20-01-2016	Second Class	EH-13945	Eric Hoffman	Consumer	United States	Los Angeles	California	90049	West	TEC-AC-1000	Technology	Accessories	Imation 8GB	90.57	3	0	11.7741
28	US-2015-150630	17-09-2015	21-09-2015	Standard Cla	TB-21520	Tracy Blumst	Consumer	United States	Philadelphia	Pennsylvania	19140	East	FUR-BO-100C	Furniture	Bookcases	Riverside Pal	3083.43	7	0.5	-1665.0522
29	US-2015-150630	17-09-2015	21-09-2015	Standard Cla	TB-21520	Tracy Blumst	Consumer	United States	Philadelphia	Pennsylvania	19140	East	OFF-BI-1000	Office Suppli	Binders	Avery Recycl	9.618	2	0.7	-7.0532
30	US-2015-150630	17-09-2015	21-09-2015	Standard Cla	TB-21520	Tracy Blumst	Consumer	United States	Philadelphia	Pennsylvania	19140	East	FUR-FU-1000	Furniture	Furnishings	Howard Mille	124.2	3	0.2	15.525
31	US-2015-150630	17-09-2015	21-09-2015	Standard Cla	TB-21520	Tracy Blumst	Consumer	United States	Philadelphia	Pennsylvania	19140	East	OFF-EN-1000	Office Suppli	Envelopes	Poly String Ti	3.264	2	0.2	1.1016
32	US-2015-150630	17-09-2015	21-09-2015	Standard Cla	TB-21520	Tracy Blumst	Consumer	United States	Philadelphia	Pennsylvania	19140	East	OFF-AR-1000	Office Suppli	Art	BOSTON Moc	86.304	6	0.2	9.7092

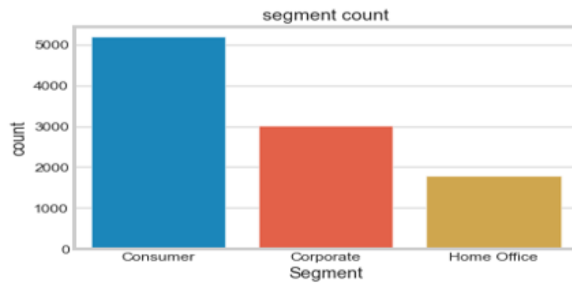
Here, we can see that the dataset has **21 columns** and total **9994 rows**.
There are several categories, such as **three(3)** in the Superstore Sales dataset and they are: **Furniture, Office Supplies and Technology**. Our aim is to analysis and forecast the sales of these different 3 categories and to see if there is any kind of time series relationship among those.

4 EXPLORATORY DATA ANALYSIS

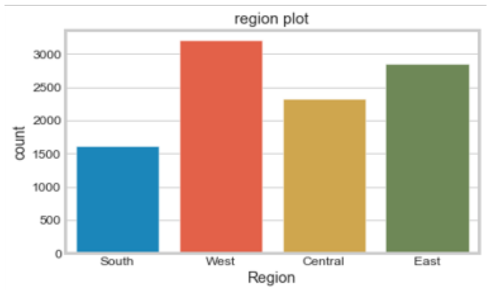
At the very beginning I have checked that if there is **any null value or missing value** in the dataset and I have found that there is **no missing value or NA** in the dataset. Then I have performed some basic **exploratory data analysis** on the dataset to gather some insightful details.



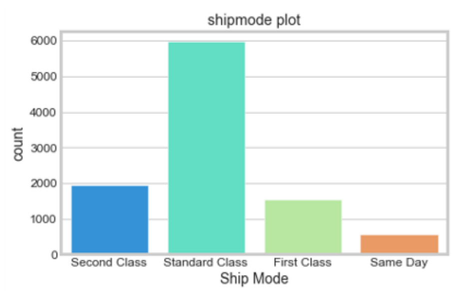
Here we can see among the **3 categories**, Office Supplies has the most count of **6026**,furniture is after that with **2121** and Technology is the last with **1847**.



Here we have **3 types of segment count**. Among them **Consumer segment** has most count, followed by **Corporate** and **Home Office** is the last.



Here the region is divided into **4 parts**. **West** has the most number of count, followed by **East**, then **Central** and **South** has the least count of all.



It can be seen Shipmode has **4 different types**. **Standard Class** has the most count, followed by **Second Class**, then **First Class** and **Same Day** shipmode has the least count of all.

From the plot it is evident that **at 0 percent discount the profit is maximum**. Then with **increase in discount, the profit decreases** and **profit is the lowest when the discount is 50 percent**. Then again, **profit starts to increase with increase in discount** and becomes **steady with higher discount**.

5 DATA PREPARATION

We have to analysis and forecast the sales of **3 categories** in the dataset.And for that we have to prepare them accordingly, so that we can work on them. We have found that **Furniture data** has timeperiod of **(2014-01-06 00:00:00 to 2017-12-30 00:00:00)**,**Office Supplies data** has timeperiod of **(2014-01-03 00:00:00 to 2017-12-30 00:00:00)**,and **Technology data** has timeperiod of **(2014-01-06 00:00:00 to 2017-12-30 00:00:00)**.
We then eliminate all the other irrelevant columns of the dataset and make **three time series** for **three different categories**.

For **Furniture Sales**, we have:

Index	**Order Date**	**Sales**
0	2014-01-06	**2573.820**
1	2014-01-07	**76.728**

For **Office Supplies Sales**, we have:

Index	**Order Date**	**Sales**
0	2014-01-03	**16.448**
1	2014-01-04	**288.06**

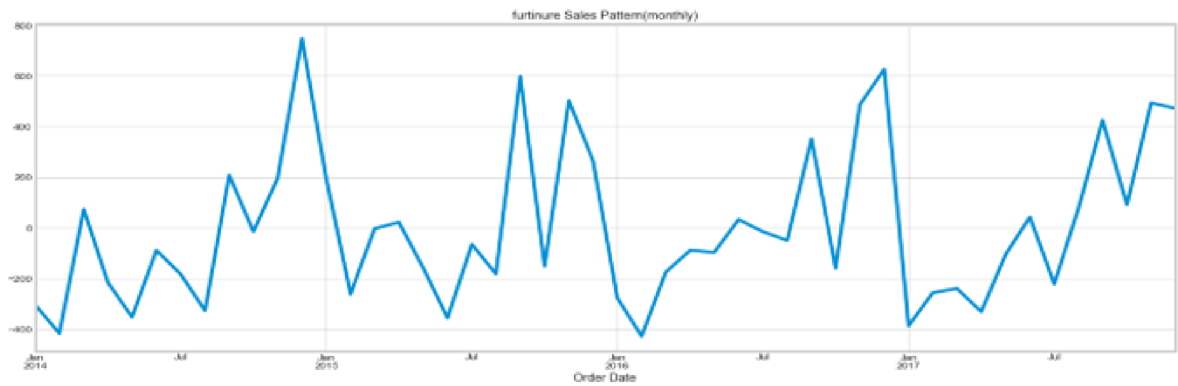
For **Technology Sales**, we have:

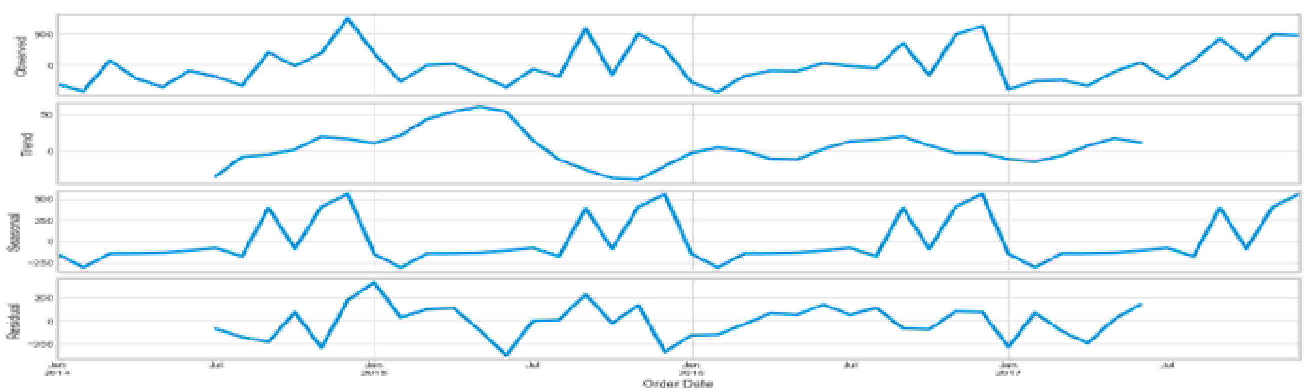
Index	**Order Date**	**Sales**
0	2014-01-06	**1147.94**
1	2014-01-09	**31.20**

6 DATA VISUALIZATION

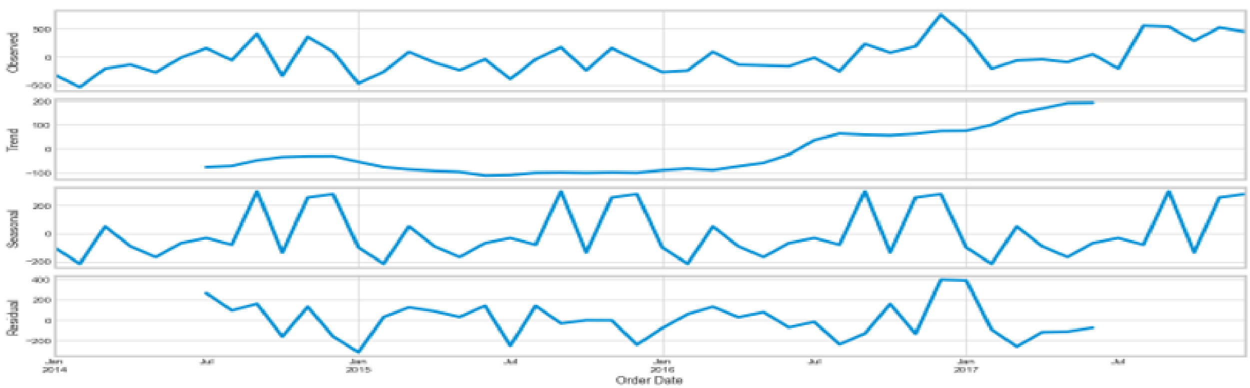
After transforming the categories into different time series dataset and **indexing over DateTime**,we have to **visualize them differently** and should try to find out some meaningful information that would help us in further analysis and forecast.

Furniture sales dataset is resampled over **month** at first and then it is **visualized** and observed its **time series pattern**.

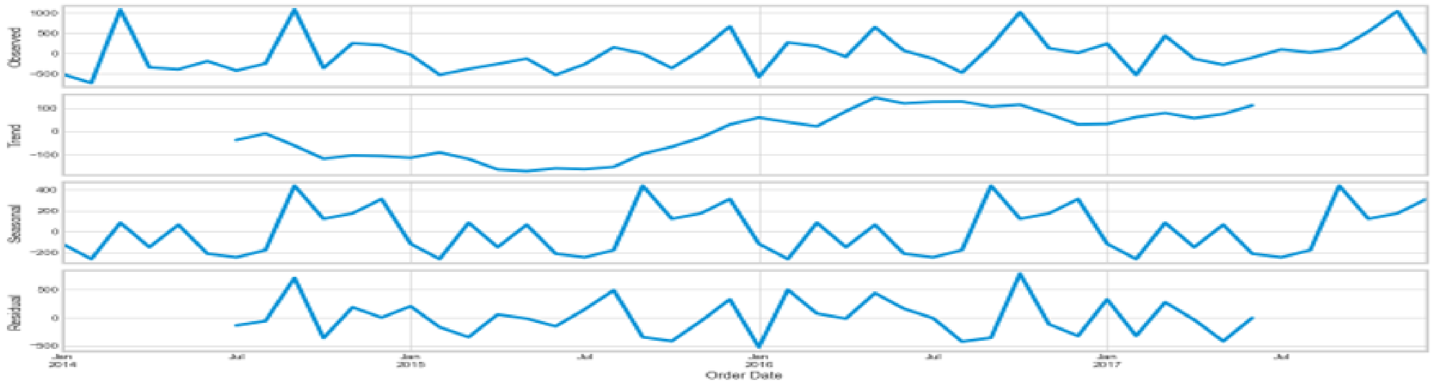




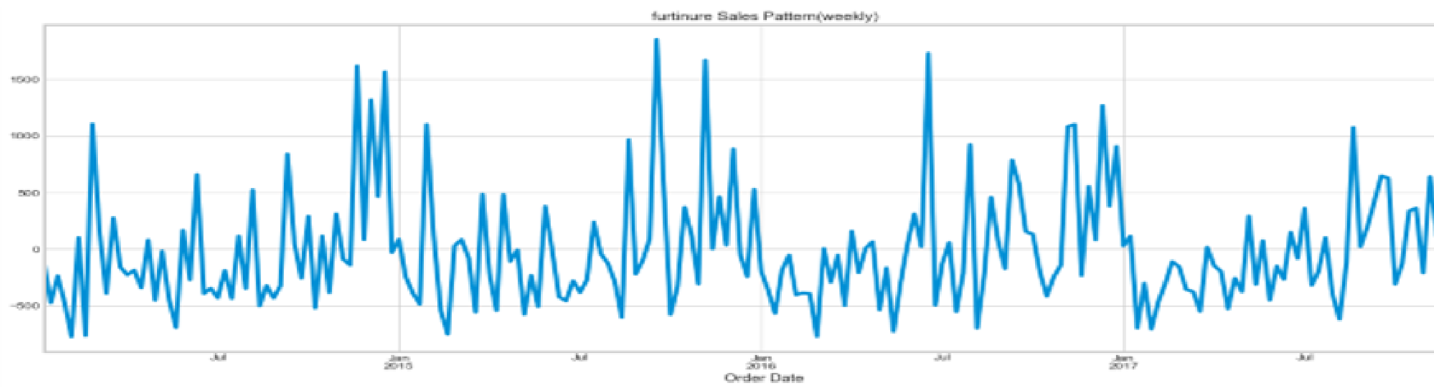
Office Supplies sales dataset is resampled over **month** at first and then it is **visualized** and observed its **time series pattern**.



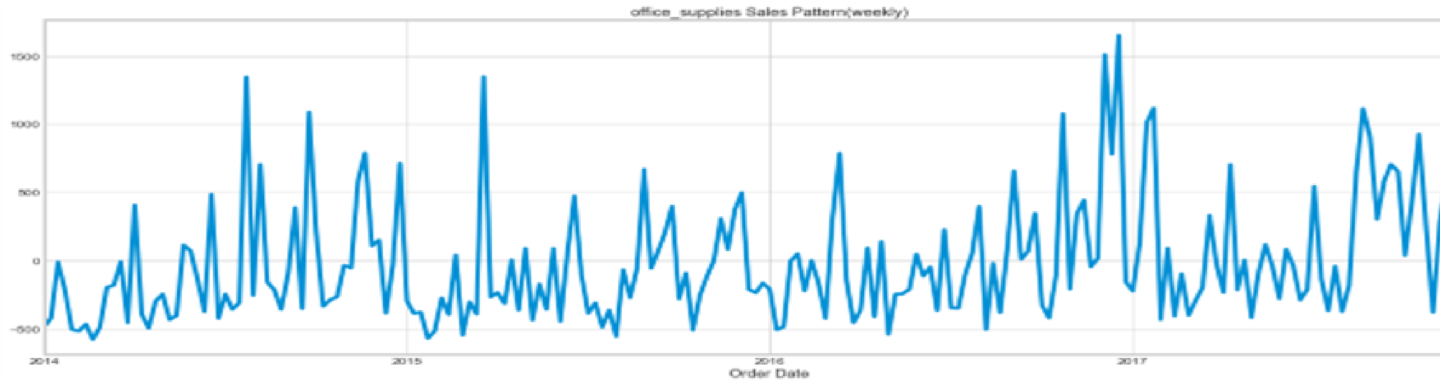
Technology sales dataset is resampled over **month** at first and then it is **visualized** and observed its **time series pattern**.

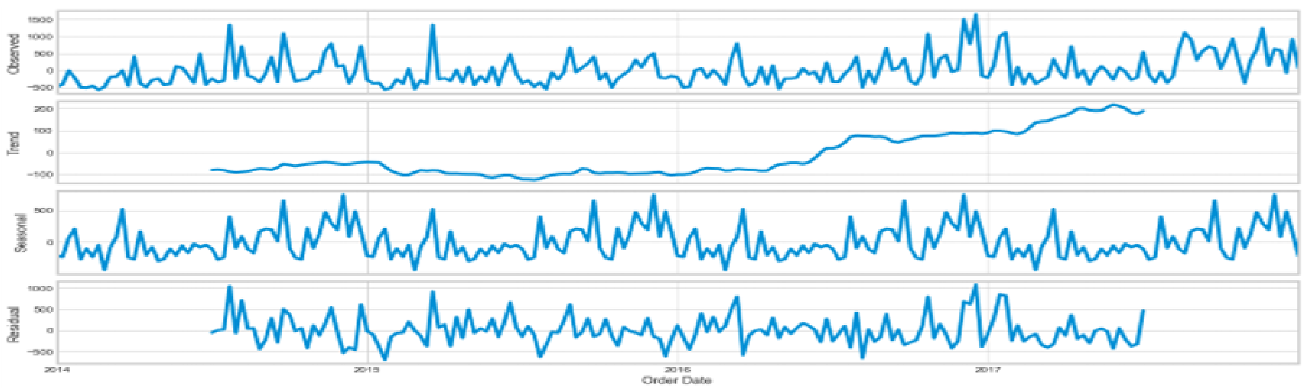


Furniture sales dataset is resampled over **week** then and it is then **visualized** and observed its **time series pattern**.

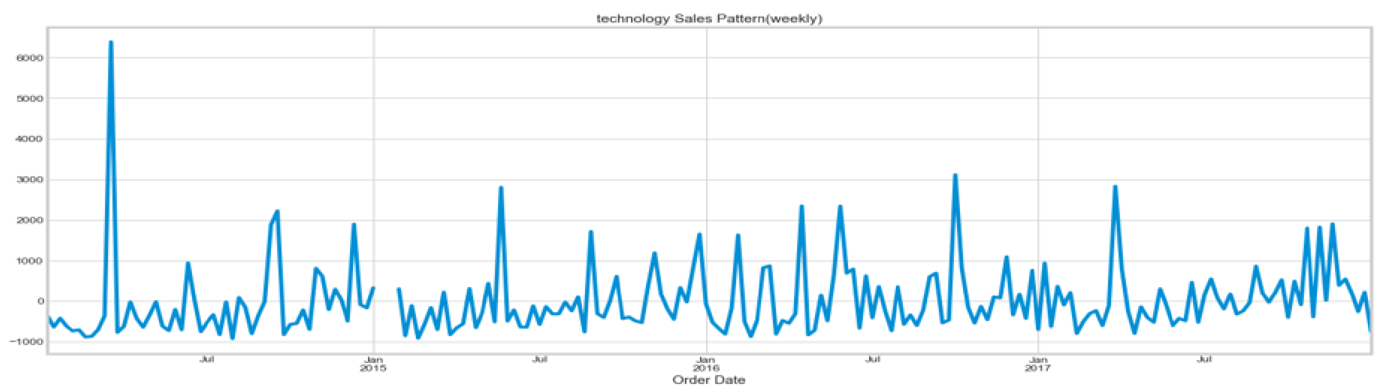


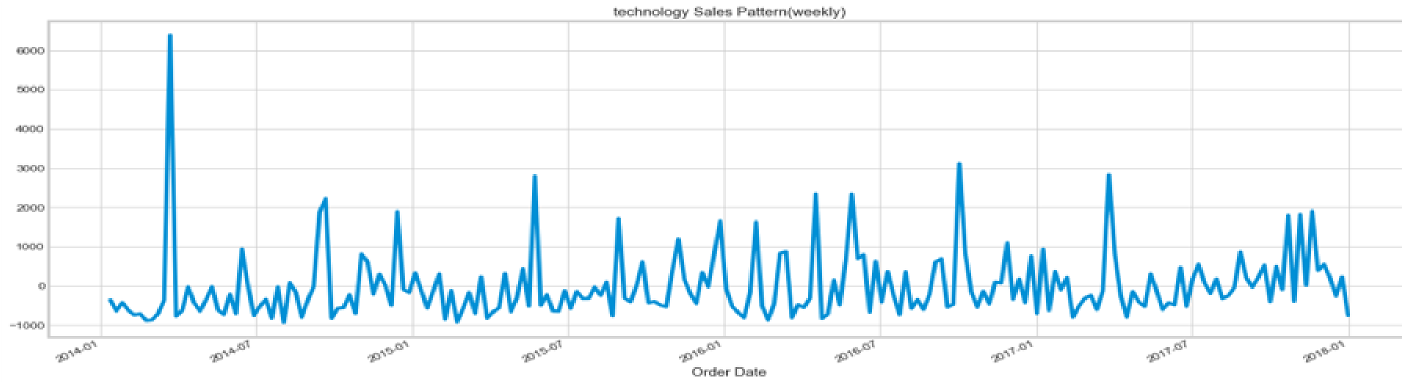
Office Supplies sales dataset is resampled over **week** then and it is then**visualized** and observed its **time series pattern**.





Technology sales dataset is resampled over **week** then and it is then **visualized** and observed its **time series pattern**.





7 STATIONARITY CHECKING

In a time series analysis, A **stationary time series** is one whose properties **do not depend on the time** at which the series is observed. Thus, **time series with trends, or with seasonality, are not stationary** — the trend and seasonality will affect the value of the time series at different times.**Summary statistics** calculated on the time series are **consistent over time**, like the mean or the variance of the observations. In statistics, the **Dickey–Fuller test** tests **the null hypothesis that a unit root is present in an autoregressive model**. The **alternative hypothesis** is different depending on which version of the test is used, but is usually **stationarity or trend-stationarity**.

H0=non-stationary process VS H1=stationary process.
If **p value greater than alpha**,fail to reject H0,i.e., **non-stationary process** and **differencing(d)** has to be done to on the dataset.
If **p value less than alpha**,reject H0,i.e., **stationary process**.

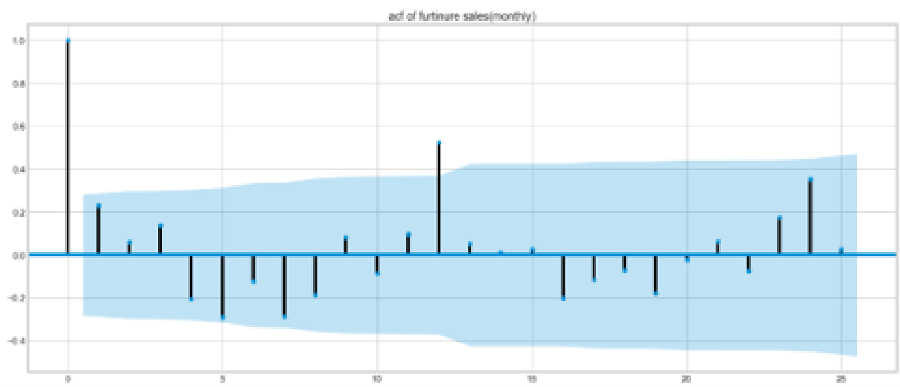
Here we have **level of significance(alpha)=0.05**.

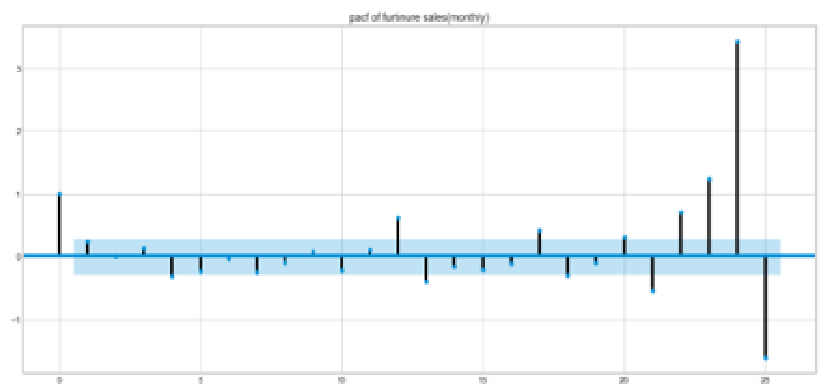
Category	p-value	Decision
Furniture(monthly)	0.00009	**stationary process**
Office supplies(monthly)	0.955	**non-stationary process**
Office supplies at 1st diff(monthly)	0.011	**stationary process**
Technology(monthly)	0.6511	**non-stationary**
Technology at 3rd diff(monthly)	0.0057	**stationary process**
Furniture(weekly)	0.0298	**stationary process**
Office supplies(weekly)	0.280	**non-stationary process**
Office supplies at 1st diff(weekly)	0.00008	**stationary process**
Technology(weekly)	0.0387	**stationary**

8 ACF PACF PLOTS

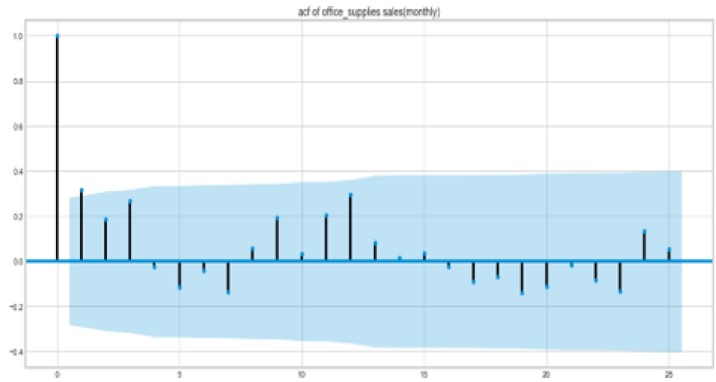
Autocorrelation Function(ACF) and **Partial Autocorrelation Function(PACF)** are two important parts in Time Series analysis.**ACF plot** is merely a bar chart of the coefficients of correlation between a time series and lags of itself.The **PACF plot** is a plot of the partial correlation coefficients between the series and lags of itself.
ACF and **PACF plots** allow one to determine the **AR** and **MA** components of an **ARIMA model**. Both the **Seasonal** and the **non-Seasonal AR** and **MA** components can be determined from the ACF and PACF plots.

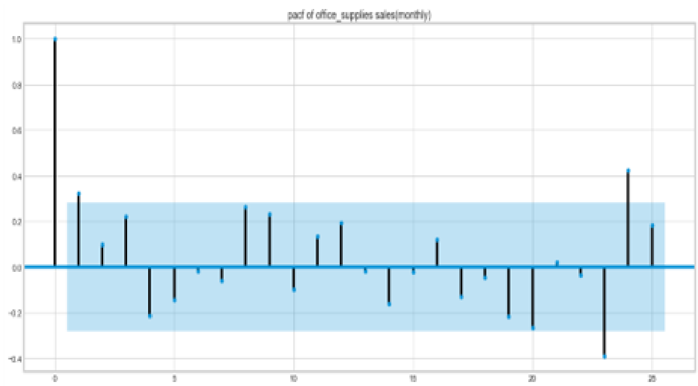
First we will see all the **acf pacf plots** of **3** different **monthly time series** data and observe their pattern.

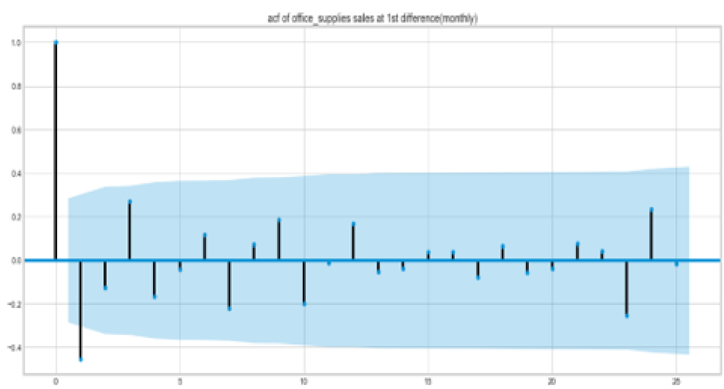


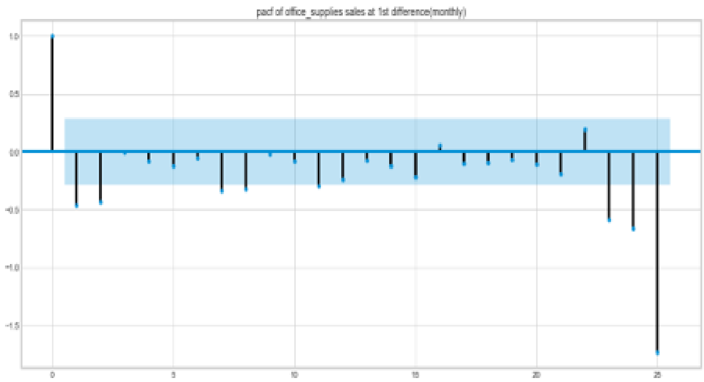


From the **acf pacf plot of Furniture sales data**, we can see that there is **order of seasonality** in the data.

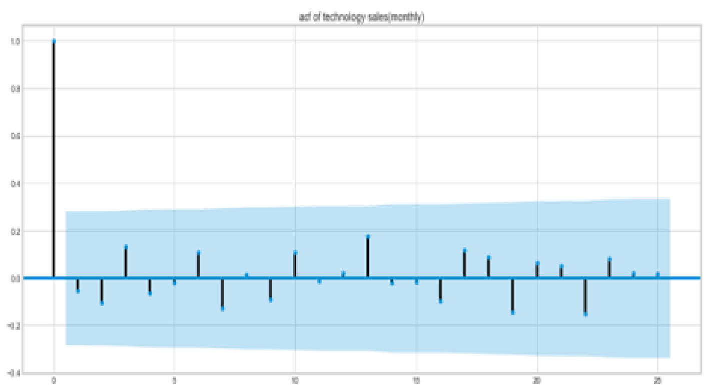

acf of office_supplies sales(monthly)

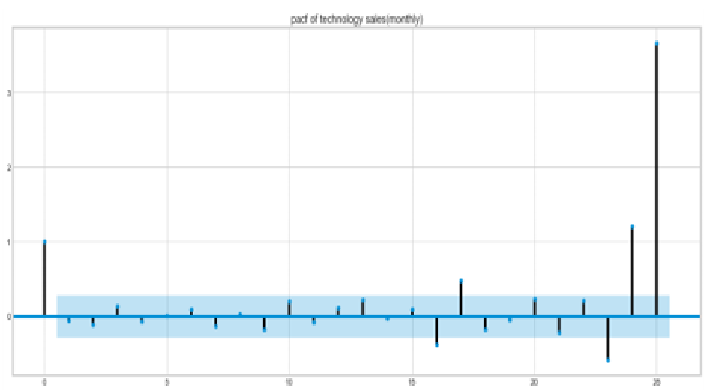

pacf of office_supplies sales(monthly)

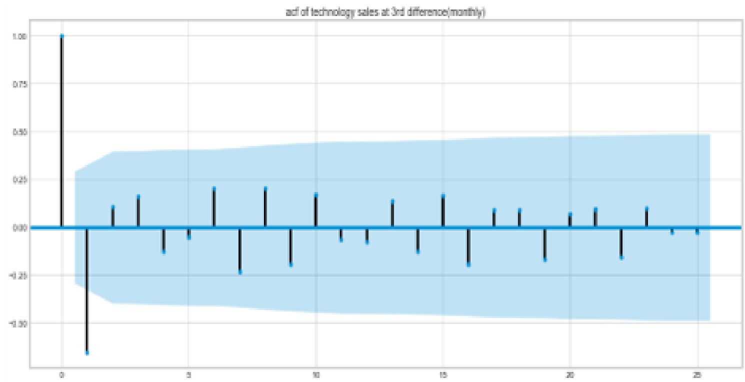

acf of office_supplies sales at 1st difference(monthly)

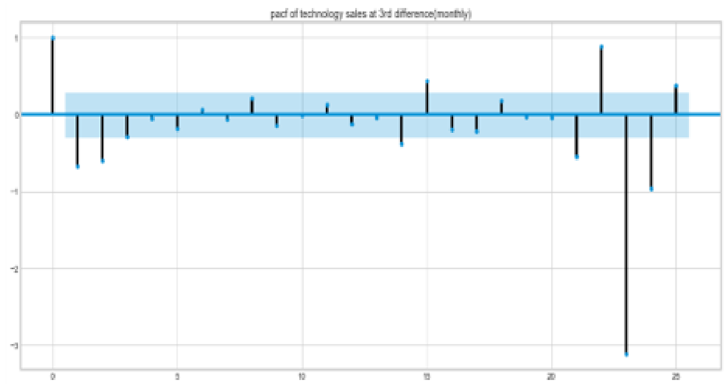


From the **acf pacf plots of Office supplies data** and of its **1st order diff**,we can see that **the trend has been eliminated but seasonality is there**.



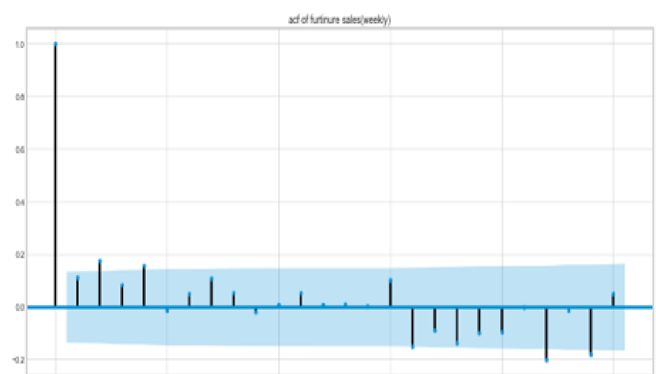


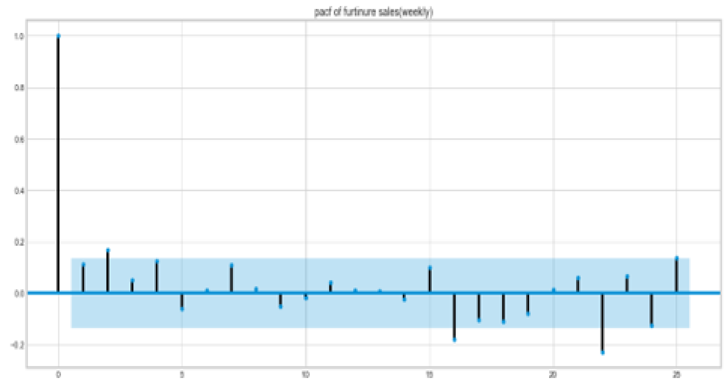




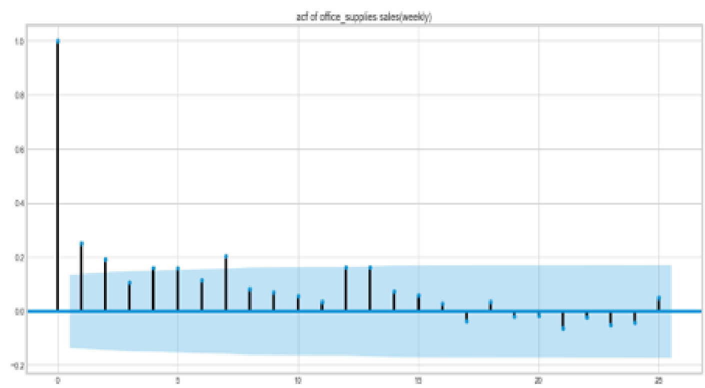
From the **acf pacf plots of Technology data** and its **3rd order diff**, we can see that **the trend has been eliminated but seasonality is there**.

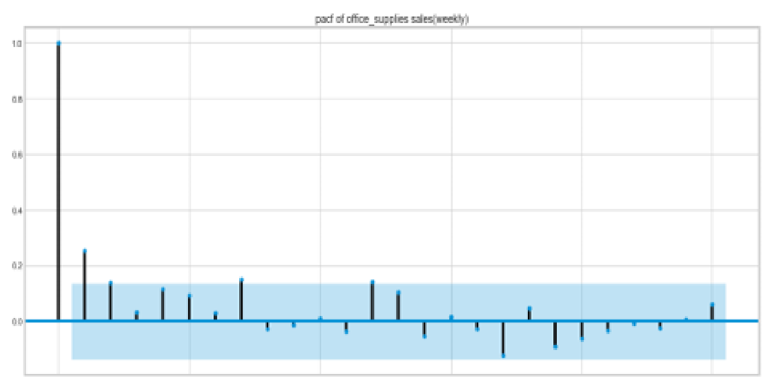
Now we will see all the **acf pacf plots** of those **3 weekly time series** data and observe their pattern.

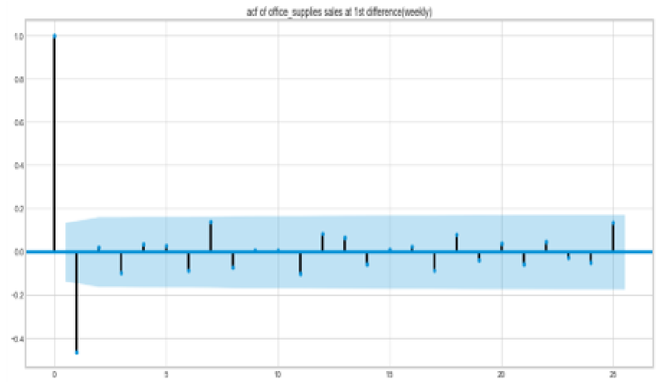


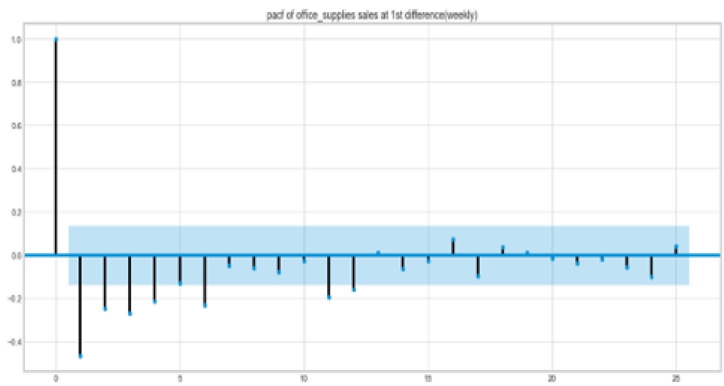


From the **acf pacf plot of Furniture sales data**, we can see that there is **order of seasonality** in the data.

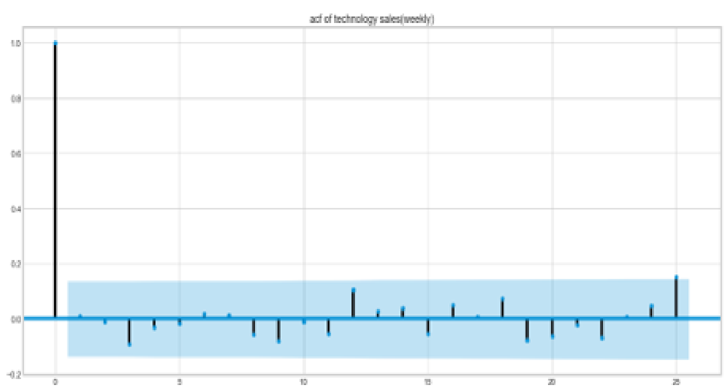


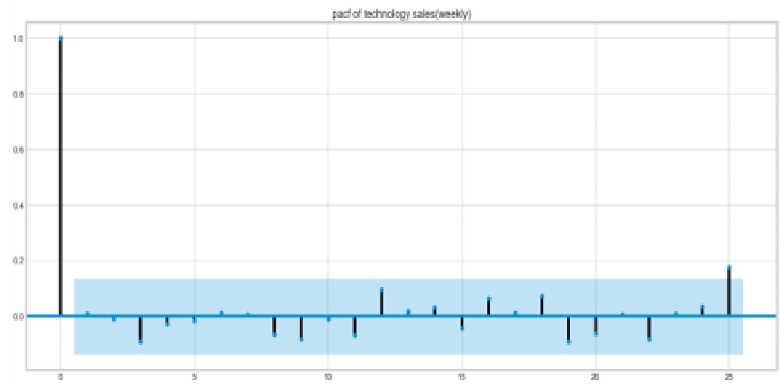






From the **acf pacf plots of Office supplies data** and of its **1st order diff**,we can see that **the trend has been eliminated but seasonality is there**.





From the **acf pacf plot of Technology sales data**, we can see that there is **order of seasonality** in the data.

9 TIME SERIES FORECASTING

Forecasting involves taking models fit on historical data and using them to predict future observations. In Time Series, there are different models for fitting and based on different scenarios and situations, we are going to apply them on our **three Time Series**, i.e., **Furniture Sales, Office Supplies Sales and Technology Sales**. And the model with the **least AIC(Akaike information criterion)**,which is **error due to prediction**, will be chosen as the **best fitted model**.

When we consider **Furniture Sales(monthly)**:

MODEL	**AIC**	**BIC**
ARMA(1,1)	690.1031	697.5879
ARIMA(1,1,1)	678.2684	685.6690
ARIMA(2,1,2)	686.2995	697.4004
ARIMA(0,1,1)*(0,1,1,12)12	**279.5568**	**288.3920**

When we consider **Furniture Sales(weekly)**:

MODEL	**AIC**	**BIC**
ARIMA(1,1,1)	3181.7677	3195.0986
ARIMA(2,1,2)	3178.6180	3198.6143
ARIMA(0,1,1)*(0,1,1,12)12	2807.8757	2816.7590

So, we can see that **Furniture Sales** time series model can best be fitted by **monthly data** and **ARIMA(0,1,1)*(0,1,1,12)12** model which has the **least AIC value** among all.

When we consider **Office Supplies Sales(monthly)**:

MODEL	**AIC**	**BIC**
ARIMA(2,1,2)	668.1807	679.2716
ARIMA(3,1,2)	667.6918	680.6429
ARIMA(0,1,1)*(0,1,1,12)12	**302.4834**	**305.6713**

When we consider **Office Supplies Sales(weekly)**:

MODEL	**AIC**	**BIC**
ARIMA(2,1,2)	3119.3551	3139.3603
ARIMA(7,1,3)	3121.2728	3161.3232
ARIMA(0,1,1)*(0,1,1,12)12	2756.6553	2778.3914

So, we can see that **Office Supplies** Sales time series model can best be fitted by **monthly data** and **ARIMA(0,1,1)*(0,1,1,12)12** model which has the **least AIC** value among all.

When we consider **Technology Sales(monthly)**:

MODEL	AIC	BIC
ARIMA(2,1,2)	719.5730	730.6739
ARIMA(3,1,2)	719.7761	732.7271
ARIMA(0,1,1)*(0,1,1,12)12	**325.4324**	**332.1786**

When we consider **Technology Sales(weekly)**:

MODEL	AIC	BIC
ARIMA(2,1,2)	3371.5322	3391.4702
ARIMA(3,1,2)	3378.4040	3401.6625
ARIMA(0,1,1)*(0,1,1,12)12	2948.1869	2979.5430

So, we can see that **Technology** Sales time series model can best be fitted by **monthly data** and **ARIMA(0,1,1)*(0,1,1,12)12** model which has the **least AIC** value among all.

10 FORECASTING RESULTS

After the **forecasting** and getting the **best fitted models** for each of the category time series model, such as **Furniture Sales**, **Office Supplies Sales** and **Technology Sales**, we have to apply those models on them and have to check how they have fared.

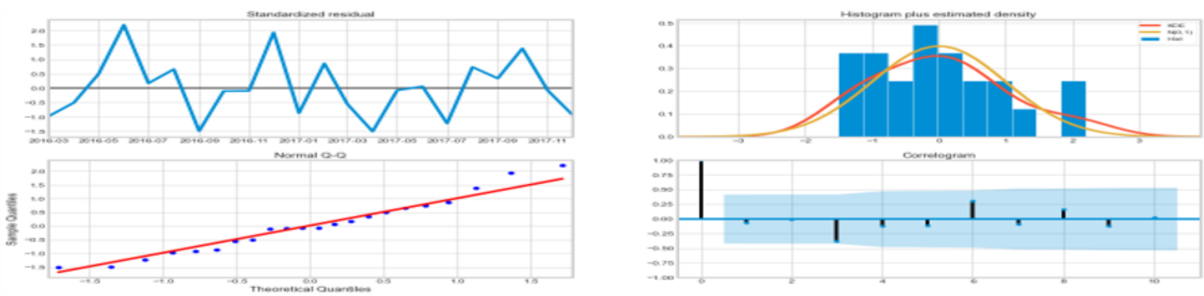
AIC value of **furniture(monthly)** is **least ARIMA(0, 1, 1)x(0, 1, 1, 12)12 - 279.55681219250735** among all:

```
In [94]: ##AIC value of furniture(monthly) is least of ARIMA(0, 1, 1)x(0, 1, 1, 12)12 - AIC:279.55681219250735 among all
         mod = sm.tsa.statespace.SARIMAX(y1_mn,
                                         order=(0, 1, 1),
                                         seasonal_order=(0, 1, 1, 12),
                                         enforce_stationarity=False,
                                         enforce_invertibility=False)

         results = mod.fit()

         print(results.summary().tables[1])

         ==============================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
         ------------------------------------------------------------------------------
         ma.L1         -1.0000   3900.897     -0.000      1.000   -7646.619    7644.619
         ma.S.L12      -3.2469      1.661     -1.954      0.051      -6.503       0.009
         sigma2      2368.6927   9.24e+06      0.000      1.000   -1.81e+07    1.81e+07
         ==============================================================================
```



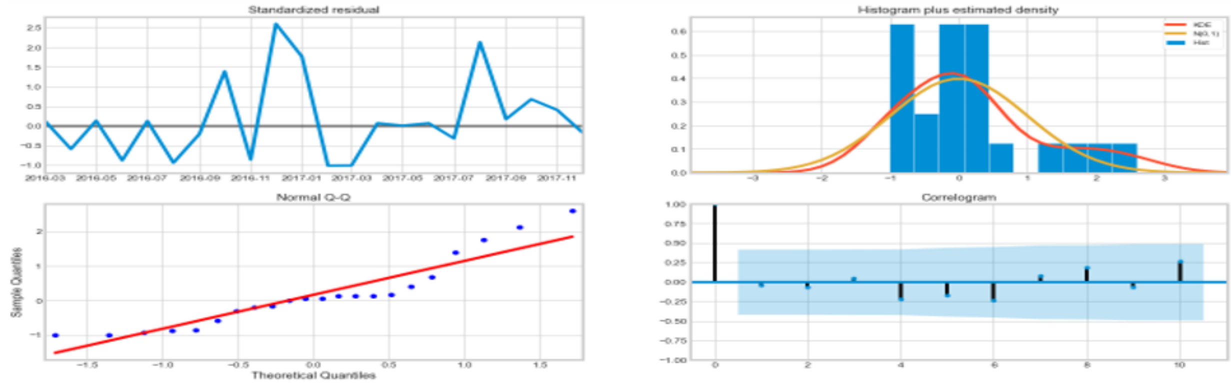
AIC value of **office supplies(monthly)** is **least ARIMA(0, 1, 1)x(0, 1, 1, 12)12 - 302.48343992251057** among all:

```
In [110]: ##AIC value of office_supplies(monthly) is least of ARIMA(0, 1, 1)x(0, 1, 1, 12)12 - AIC:302.48343992251057 among all
          mod2 = sm.tsa.statespace.SARIMAX(y2_mn,
                                           order=(0, 1, 1),
                                           seasonal_order=(0, 1, 1, 12),
                                           enforce_stationarity=False,
                                           enforce_invertibility=False)

          results2 = mod2.fit()

          print(results2.summary().tables[1])

          ==============================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
          ------------------------------------------------------------------------------
          ma.L1         -0.8701      0.218     -3.983      0.000      -1.298      -0.442
          ma.S.L12      -1.0079     53.047     -0.019      0.985    -104.979     102.963
          sigma2      5.591e+04   2.98e+06      0.019      0.985   -5.78e+06    5.89e+06
          ==============================================================================
```



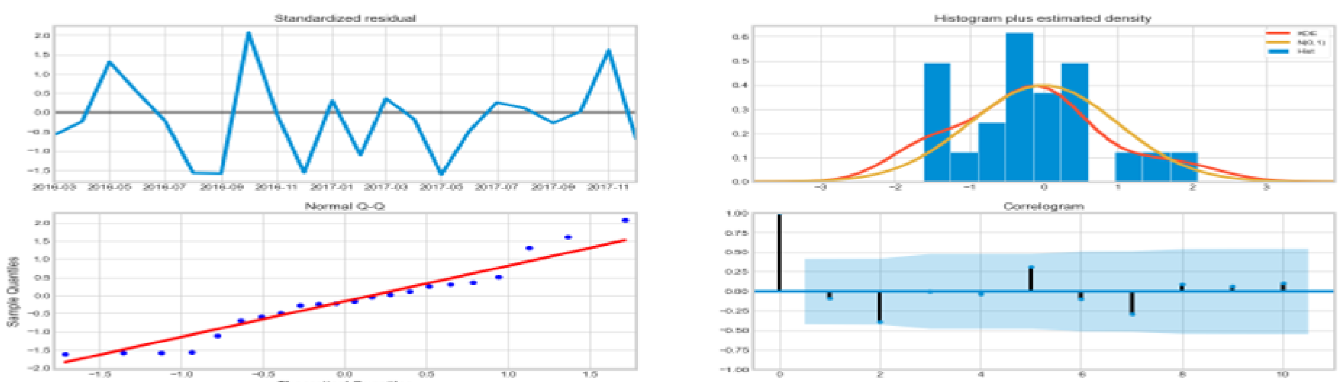
AIC value of **technology(monthly)** is **least ARIMA(0, 1, 1)x(0, 1, 1, 12)12 - 325.43244967810364** among all:

```
In [120]: ##AIC value of technology(monthly) is least of ARIMA(0, 1, 1)x(0, 1, 1, 12)12 - AIC:325.43244967810364 among all
          mod3 = sm.tsa.statespace.SARIMAX(y3_mn,
                                          order=(0, 1, 1),
                                          seasonal_order=(0, 1, 1, 12),
                                          enforce_stationarity=False,
                                          enforce_invertibility=False)

          results3 = mod3.fit()

          print(results3.summary().tables[1])
```

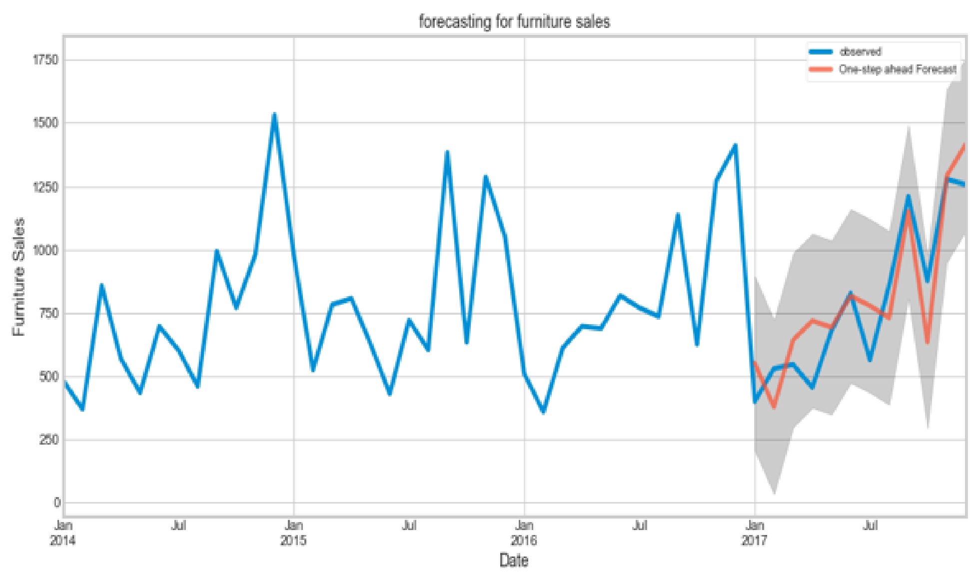
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -1.0002      5.944     -0.168      0.866     -12.650      10.650
ma.S.L12      -1.0434      5.784     -0.180      0.857     -12.380      10.294
sigma2      1.568e+05   3.66e-05   4.28e+09      0.000    1.57e+05    1.57e+05
==============================================================================
```



11 VALIDATING FORECASTING RESULTS

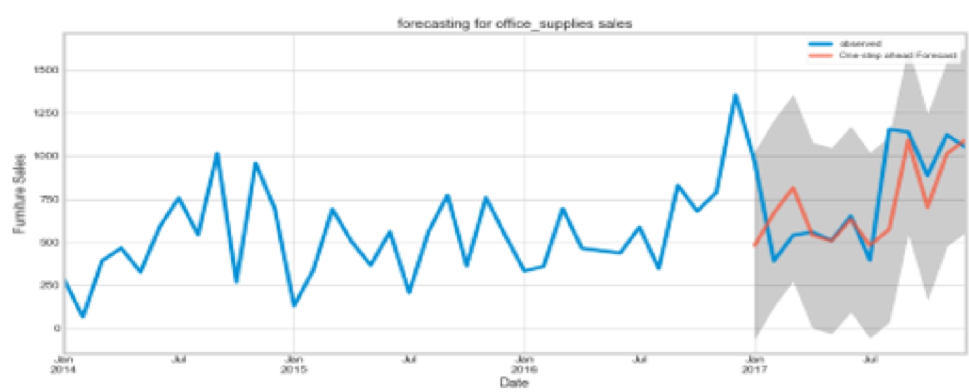
After getting fitted models and results for all three category time series model **Furniture Sales**,**Office Supplies Sales** and **Technology Sales**, we have to **verify how good our fitted model works** on real life scenario.
For the **validation of results**,we will use two measure of statistics,**Mean Squared Error(MSE)** and **Root Mean Squared Error(RMSE)**.MSE is the **mean of square of deviation from actual to prediction data** and RMSE is the **square root of MSE**. **Lesser the RMSE,Better the model prediction**.

Forecasting for **furniture sales**:



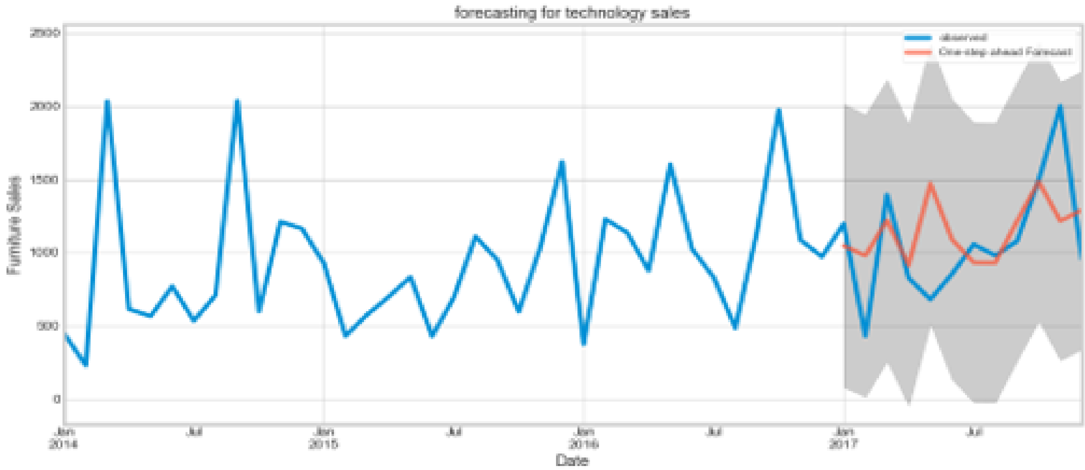
The Mean Squared Error of furniture forecasts is 22993.5664.
The Root Mean Squared Error of furniture forecasts is 151.64.

Forecasting for **office supplies sales**:



The Mean Squared Error of furniture forecasts is 65844.6001.
The Root Mean Squared Error of furniture forecasts is 256.6.

Forecasting for **technology sales**:



The Mean Squared Error of furniture forecasts is 136874.9803.
The Root Mean Squared Error of furniture forecasts is 369.97.

12 VISUALIZING FORECASTS

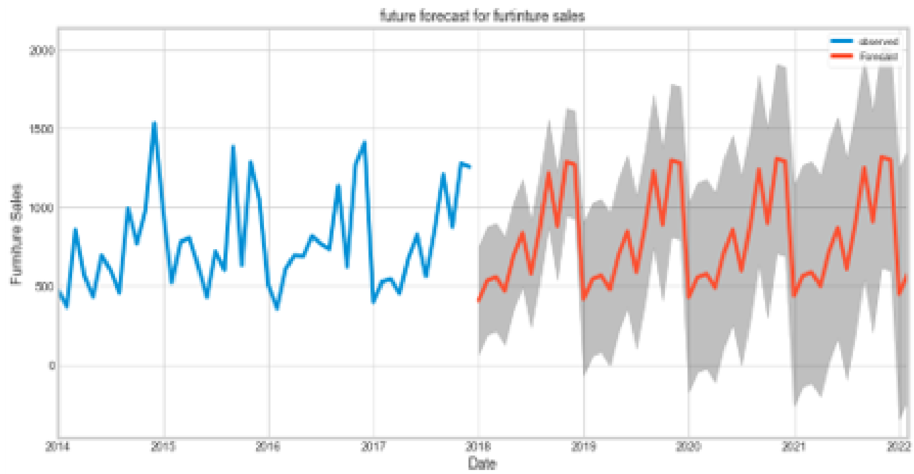
I am **visualizing forecasts** for **Furniture Sales** in next **50 steps** and observe how it looks like.

visualizing forecasts

In [134]:
```
pred_uc1 = results.get_forecast(steps=50)
pred_ci1 = pred_uc1.conf_int()

ax = y1_mn.plot(label='observed', figsize=(14, 7))
pred_uc1.predicted_mean.plot(ax=ax, label='Forecast')
ax.fill_between(pred_ci1.index,
                pred_ci1.iloc[:, 0],
                pred_ci1.iloc[:, 1], color='k', alpha=.25)
ax.set_xlabel('Date')
ax.set_ylabel('Furniture Sales')

plt.legend()
plt.title("future forecast for furtinture sales")
plt.show()
```

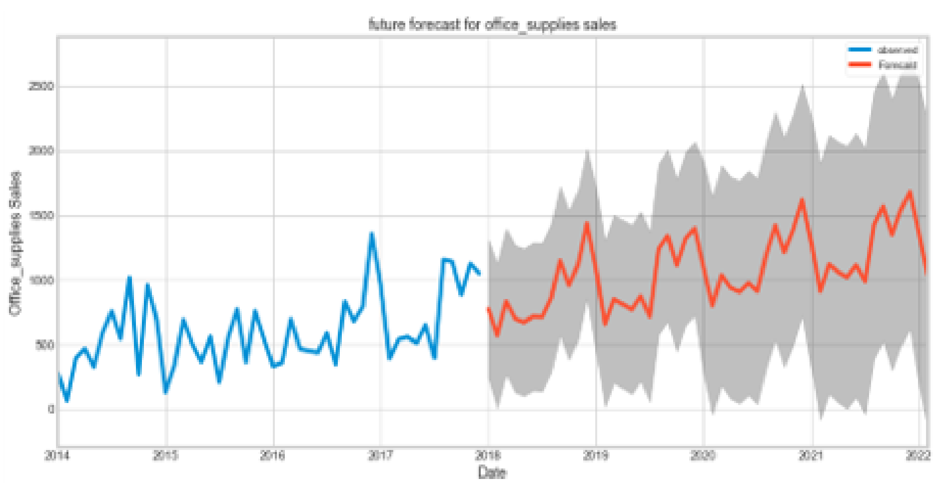


I am **visualizing forecasts** for **Office supplies Sales** in next **50 steps** and observe how it looks like.

```
In [135]: pred_uc2 = results2.get_forecast(steps=50)
          pred_ci2 = pred_uc2.conf_int()

          ax = y2_mn.plot(label='observed', figsize=(14, 7))
          pred_uc2.predicted_mean.plot(ax=ax, label='Forecast')
          ax.fill_between(pred_ci2.index,
                          pred_ci2.iloc[:, 0],
                          pred_ci2.iloc[:, 1], color='k', alpha=.25)
          ax.set_xlabel('Date')
          ax.set_ylabel('Office_supplies Sales')

          plt.legend()
          plt.title("future forecast for office_supplies sales")
          plt.show()
```

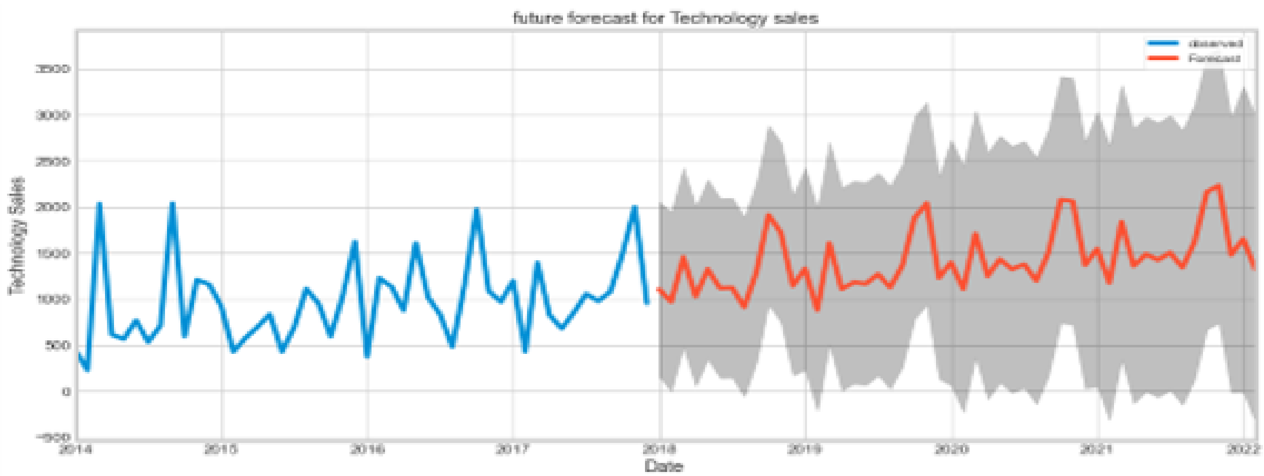


I am **visualizing forecasts** for **Technology Sales** in next **50 steps** and observe how it looks like.

```
In [136]: pred_uc3 = results3.get_forecast(steps=50)
          pred_ci3 = pred_uc3.conf_int()

          ax = y3_mn.plot(label='observed', figsize=(14, 7))
          pred_uc3.predicted_mean.plot(ax=ax, label='Forecast')
          ax.fill_between(pred_ci3.index,
                          pred_ci3.iloc[:, 0],
                          pred_ci3.iloc[:, 1], color='k', alpha=.25)
          ax.set_xlabel('Date')
          ax.set_ylabel('Technology Sales')

          plt.legend()
          plt.title("future forecast for Technology sales")
          plt.show()
```



13 CAUSALITY CHECKING OF TIME SERIES

Causality concerns relationships where **a change in one variable necessarily results in a change in another variable**.There are **three conditions for causality**: **covariation**, **temporal precedence**, and **control for "third variables"**. The latter comprise alternative explanations for the observed causal relationship.
The **Granger causality test** is a statistical hypothesis test for determining whether one time series is useful for forecasting another. If probability value is less than any level, then the hypothesis would be rejected at that level.**Granger causality** is a statistical concept of causality that is based on prediction. According to Granger causality,**if a signal X1 "Granger-causes" (or "G-causes") a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone**.
When time series X Granger-causes time series Y, the patterns in X are approximately repeated in Y after some time lag. Thus, past values of X can be used for the prediction of future values of Y. **A time series X is said to Granger-cause Y if it can be shown, usually through a series of t-tests and F-tests on lagged values of X (and with lagged values of Y also included), that those X values provide statistically significant information about future values of Y.**
H0: Xt does not granger causes Yt VS H1: Xt granger causes Yt.

If **p value greater than alpha**,fail to reject H0,i.e.,**not have granger causes**.If **p value less than alpha**,reject H0,i.e.,**have granger causes**.
Granger Causality Test is one-way test. Here ,we test on X variable granger causes Y but not the other way round. And so, for that we should keep in mind the order we put random variables in.**The first one put, will be Y, dependent variable and second one put, will be X, independent variable, which will granger-cause.**

The **first thing,we have to do for causality checking is to see if obsereved time series are of same time span or not**.If not,we have to make them in same time span. Here,**Furniture sales** and **Technology sales** are of **same time span** but **Office supplies sales** is **not**. So, we have **to make it into same of other two** and then we can **conduct the causality checking**.
For **Causality checking**,we have to **select no of lags** for which we **want to test**.We have selected **no of lags as 5** and **alpha=0.05**.

casuality checking

In [137]:
```
office_supplies = office_supplies.iloc[3:]
office_supplies.head()
```

Out[137]:

Order Date	Sales
2014-01-06	685.340
2014-01-07	10.430
2014-01-09	9.344
2014-01-10	2.890
2014-01-13	2027.116

granger causality checking when y=furniture,x=office supplies:
p Values per lag - [0.2951, 0.6339, 0.5935, 0.7762, 0.6253]
granger causality checking when y=furniture,x=technology:
p Values per lag - [0.2626, 0.5257, 0.5133, 0.5408, 0.4203]
We can see that **p-values for both tests at all the lag 5** is **more than alpha** and so we can conclude that **for forecasting of Furniture sales, Office supplies sales and Technology sales is not dependent** and **Furniture sales is not granger caused by Office supplies sales and Technology sales.**

granger causality checking when y=office supplies,x=furniture:
p Values per lag - [0.7878, 0.8714, 0.9486, 0.9751, 0.9263]
granger causality checking when y=office supplies,x=technology:
p Values per lag - [0.998, 0.7859, 0.8105, 0.81, 0.6444]
We can see that **p-values for both tests at all the lag 5** is **more than alpha** and so we can conclude that **for forecasting of Office supplies sales, Furniture sales and Technology sales is not dependent** and **Office supplies sales is not granger caused by Furniture sales and Technology sales.**

granger causality checking when y=technology,x=furniture:
p Values per lag - [0.4353, 0.6432, 0.762, 0.7089, 0.8634]
granger causality checking when y=technology,x=office supplies:
p Values per lag - [0.3428, 0.1854, 0.1117, 0.167, 0.3111]
We can see that **p-values for both tests at all the lag 5** is **more than alpha** and so we can conclude that **for forecasting of Technology sales, Furniture sales and Office supplies sales is not dependent** and **Technology sales is not granger caused by Furniture sales and Office supplies sales.**
So at the end, we can conclude that **whatever forecasting we have done for the different time series analysis is sufficient and properly applicable**.