REGRESSION TIME SERIES MODEL COURSE
GROUP PROJECT

Gaurav Patidar	**20AG3FP30**
Alankrita Roy	**20AG3FP15**
Abhijeet Jena	**20AG3FP48**
Sparsha Sherke	**20AG3FP41**

INDEX

Definition of Linear Regression Analysis

A model called linear regression analysis, in which one variable predicts the response variable, aims to explain the functional relationship between two variables. Regression analysis often has three goals: to assess the strength of the association, to evaluate the impact, and to forecast or predict

Basic Model of Simple Linear Regression

$$Y = \alpha + \beta x + \varepsilon$$

Where:

Y is the dependent / response / dependent variable

X is independent variables / predictors / independent α is the intercept parameter (constant)
β slope parameter (coefficient)

ε is a residual which is a random

A simple linear regression model is an equation that states the relationship between one predictor variable (X) and one response variable (Y), which is usually depicted in a straight line. Regression analysis is a model that attempts to explain the functional relationship between two variables, where one variable acts as a predictor of the response variable
Y = a + bX

a = constant

b = regression coefficient

Y = dependent variable / dependent variable / dependent variable (incident)

X = independent variable / independent variable / variable predictor (cause)

1. **Simple Linear Regression Testing Steps**

This test is conducted to determine whether the independent variable (x) affects the dependent variable (Y). Hypothesis testing of the regression coefficient is carried out through the following steps(statistics, 2020):

a) Significance Test of Constants a

- Hypothesis:

- $Ho = \alpha = 0$ →constanta has no significant effect
- $H1 = \alpha \neq 0$ →constantsta has a significant effect

- Determine the significance level of α {find t table with df = n-2}
- Test Statistics: t = = depending on the initial hypothesis Test = $(a-\alpha) sx\sqrt{n}(n-1)$
$Se\sqrt{\sum xi}$ 2

- Test criteria:
if $|thitung| \geq ttabel$ then Ho is rejected
If $|thitung| < ttabel$ then Ho is accepted

b) Test of Significance of Coefficient b

- Hypothesis:

- $Ho = \beta = 0$ →Coefficient has no significant effect
- $H1 = \beta \neq 0$ → Coefficient has a significant effect

- Determine the significance level of α {find t table with df = n-2}

- Test Statistics: t = : t = $(b-\beta 0)sx\sqrt{(n-1)}\,Se$, $\beta 0$= depending on the initial hypothesis

- Test criteria:
if $|thitung| \geq ttabel$ then Ho is rejected
If $|thitung| < ttabel$ then Ho is accept

2. **Correlation Coefficient and Coefficient of Determination**

Correlation coefficient is used to measure the degree of closeness of the relationship between the independent variable and the dependent variable. The coefficient of determination is used to measure how much influence the independent variable has on the dependent variable. The measurement can be used the Pearson correlation formula or the formula below:

The Pearson Correlation Coefficient formula is as follows:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y}^2 - (\sum \mathbf{y})^2]}}$$

LEAST SQUARES AND THE FITTED MODEL

After determining the model, the next step is to determine the method for estimating the parameters or regression coefficients from the formed model. The regression coefficient is a parameter and its value is unknown, but these parameters can be estimated from the sample data. We will use least squares method for estimation .

Ordinary Least Squares (OLS) Method The least squares method is a method for determining the estimation linear equation by selecting one linear curve from several possible linear curves that can be made from existing data that has the smallest error from the actual data with the estimation data. If simple linear regression equation: Y=a+bX

The regression coefficients can be calculated by the formula:

$$\boxed{a = \frac{\sum y_i - b\sum x_i}{n}} \qquad \boxed{b = \frac{n.\sum xi.yi - (\sum xi).(\sum yi)}{n.(\sum x_i^2) - (\sum xi)^2}}$$

Sample case

There was a sample of 12 students of Class IX at SMPN 1 Medan which were taken for the test scores and final scores for each student. We want to see how the test scores affect the final grade . As for the following is the data:

No.	Test score (x)	Final Grade (y)
1.	65	85
2.	50	74
3.	55	76
4.	65	90
5.	55	85
6.	70	87
7.	65	94
8.	70	98
9.	55	81
10.	70	91
11.	50	76
12.	55	74

STEPS:

1. Write down the regression equation, then explain what the equation means!
2. Describe the linearity of the data, then conclude!
3. The partial significance test for the constant value and the regression coefficient? Use a significance level of 0.05 (5%).
4. Simultaneous significance test of the parameter assessment results? Use the real level 0.05 (5%)
5. How strong is the relationship between test scores and final grades?
6. What is the coefficient of determination? Explain what that value means!

ANSWER!

1. Its Regression Equation and Its Interpretation

No.	Test Score (x)	Final grade (y)	x^2	y^2	$x * y$
1.	65	85	4,225	7,225	5,525
2.	50	74	2,500	5,476	3,700
3.	55	76	3,025	5,776	4,180
4.	65	90	4,225	8,100	5,850
5.	55	85	3,025	7,225	4,675
6.	70	87	4,900	7,569	6,090
7.	65	94	4,225	8,836	6.110
8.	70	98	4,900	9,604	6,860
9.	55	81	3,025	6,561	4,455
10.	70	91	4,900	8,281	6,370
11.	50	76	2,500	5,776	3,800
12.	55	74	3,025	5,476	4,070
amount	725	1,011	44,475	85,905	61,685

$\sum x_i = 725$ $\qquad$ $\sum x_i\, y_i = 61.685$

$\sum y_i = 1011$ $\qquad$ $\bar{x} = \frac{\sum x_i}{n} = \frac{725}{12} = 60{,}42$

$\sum x_i^2 = 44.475$ $\qquad$ $\bar{y} = \frac{\sum y_i}{n} = \frac{1011}{12} = 84{,}25$

$\sum y_i^2 = 85.905$

$$b = \frac{n.\sum xi.yi - (\sum xi).(\sum yi)}{n.(\sum x^2 i) - (\sum xi)^2} = \frac{(12).(61.685) - (725)(1.011)}{(12).(44.475) - (725)^2} \frac{(740.220) - (732.975)}{(533.700) - (525.625)} = \frac{7.245}{8.075} = 0{,}897$$

$$a = \frac{\sum y_i - b \sum x_i}{n} = \frac{84{,}25 - (0{,}897).(60{,}42)}{12} = 30{,}053$$

Then the regression equation is obtained:

Y = a + bx

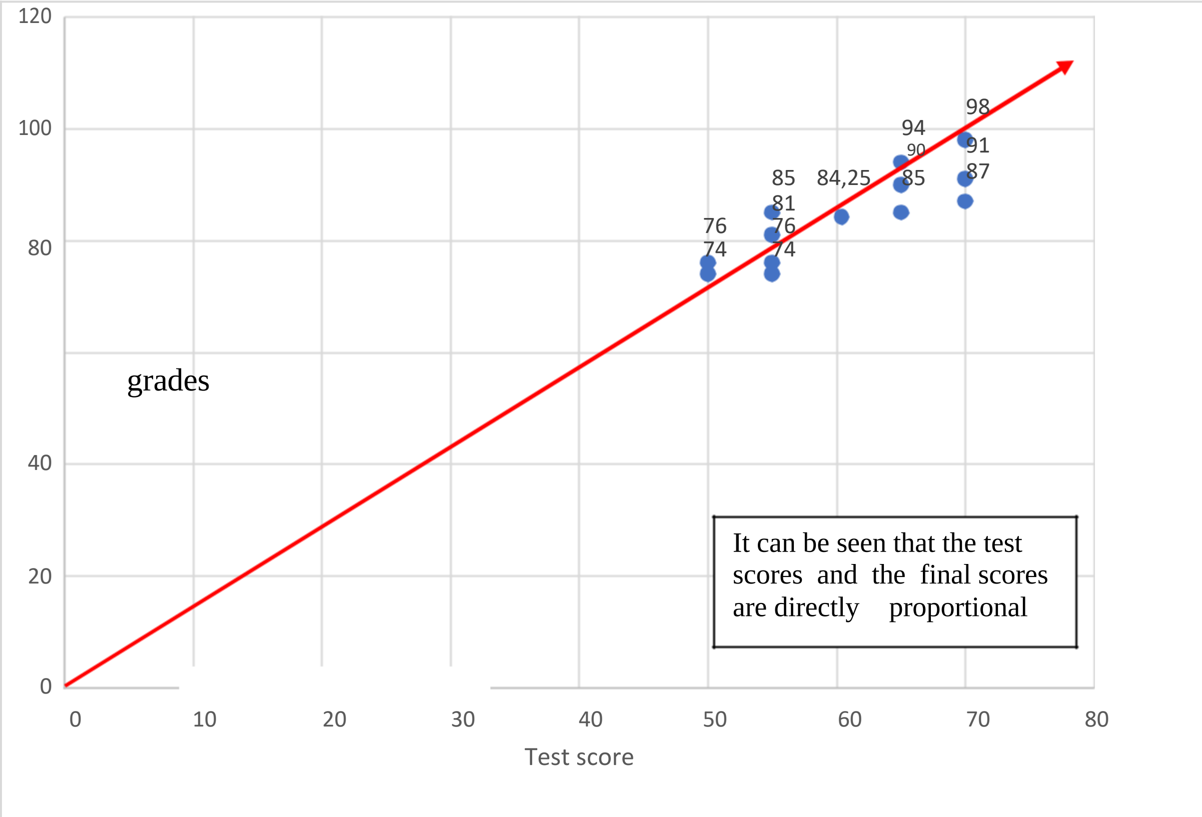
Y = 30.053 + 0.897x

From the above equation, it can be interpreted:

constant a = 30.053→ Expresses the average final score of students on the exam (X) zero amounted to **30,053**

coefficient b = 0.897→ Every increase of 1 in the test score (X), then the grade end (Y) students are likely to increase amounting to 0.897

2. Scatter Plot Data linearity



3. Partial Significance Test Parameters constants a and coefficient b

$\sum x_i = 725$ $\sum x_i^2 = 44,475$

$\sum y_i = 1011$ $\sum y_i^2 = 85,905$

$\sum x_i\, y_i = 61,685$ $n = 12$

$s_x{}^2 = \frac{n.\sum xi^2-(\sum xi)^2}{n.\,(n-1)} = \frac{(12).(44.475)-(725)^2}{12.(12-1)} = \frac{533.700-525.625}{12.(11)} = \frac{8.075}{132}$ = 61,174

$s_x = \sqrt{61{,}174}$ = 7,821

$s_y{}^2 = \frac{n.\sum yi^2-(\sum yi)^2}{n.\,(n-1)} = \frac{(12).(85.905)-(1011)^2}{12.(12-1)} = \frac{1.030.860-1.022.121}{12.(11)} = \frac{8.739}{132}$ = 66,205

$s_y = \sqrt{66{,}205}$ = 8,137

$S_e{}^2 = \frac{n-1}{n-2}(s_y{}^2 - b^2.s_x{}^2) = \frac{12-1}{(12-2)}$.[66.205-(0,897^2).(61.174)]=$\frac{11}{10}$.[66.205-(0,805)(61.174)]

=(1,1).(66.205-49.245)=(11).(16,96)= 18,66

$S_e = \sqrt{18{,}66}$ = 4,319

a) <u>Significance test of constants a</u>

1) Hypothesis:

$H_o = \alpha = 0 \rightarrow$ constantsta has no significant effect on the final value

$H_1 = \alpha \neq 0 \rightarrow$ constant has a significant effect on the final value

2) Significance Level

α= 5% = 0,05/2 = 0,025→ see Tabel t

df=n-2 =12-2=10 → $t\ tabel$ = 2,228

Test Statistics

$$t = \frac{(a-\alpha)s_x\sqrt{n(n-1)}}{S_e\sqrt{\sum x_i{}^2}} = \frac{(30{,}053-0).(7{,}821)\sqrt{12(12-1)}}{(4{,}319)\sqrt{44.475}} = \frac{(30{,}053).(7{,}821)\sqrt{132}}{(4{,}319).(210{,}891)}$$

$$= \frac{(30{,}053).(7{,}821).(11{,}489)}{(910{,}838)} = \frac{2700{,}426}{(910{,}838)} = 2{,}965 \rightarrow t_{hitung}$$

3) Test Criteria

If, $|t_{hitung}| \geq t_{tabel}$ then H_o it is rejected

If, $|t_{hitung}| < t_{tabel}$ then H_o accepted

4) **Conclusion:**

Because 2,965 > 2,228 → then Ho is rejected

"$\boldsymbol{H_1}$" → Constants have a significant effect on the final value

b) <u>Significance coefficient test b</u>

1) Hypothesis:

H_o= β= 0 → the coefficient of test scores has no significant effect on the final score

H_1=β ≠ 0 → the coefficient of the test score has a significant effect on the final score

2) Significance Level

α = 5% = 0,05/2 = 0,025→ see Tabel t

Df = n-2 = 12 – 2 = 10 → $t\ tabel$ = 2,228

3) Test Statistics

$$t = \frac{(b-\beta_0)s_x\sqrt{(n-1)}}{S_e} = \frac{(0{,}897-0).(7{,}821)\sqrt{(12-1)}}{(4{,}319)} = \frac{(0{,}897).(7{,}821).(3{,}317)}{(4{,}319)} = \frac{(23{,}270)}{(4{,}319)} = 5{,}388 \rightarrow t_{hitung}$$

4) Test Criteria

If $|t_{hitung}| \geq t_{tabel}$ then H_o is rejected

If $|t_{hitung}| < t_{tabel}$ then H_o accepted

5) **Conclusion:**

Because 5,388 > 2,228 → then Ho is rejected

"H_1" → The test score coefficient has a significant effect on the final score

4. <u>Silmultan Significance Test Results Of Parameter Assessment</u>

- Hypothesis:

H_o = independent has no effect on the dependent variable

H_1 = independent variabel have effects to the dependent variable

- Significance level

α= 5% = 0,05 → see tabel F

df_1 =1; df_2 =12-2=10 → $F\ tabel$ = 4,96

- Test Statistics ($\sum y_i$= 1011 dan $\sum y_i^2$= 85.905)

▪ JKT = $\sum {y_i}^2 - \frac{(\sum y_i)^2}{n}$ = (85.905) $- \frac{(1011)^2}{12}$ = (85.905) $- \frac{(1.022.121)}{12}$

= 85.905-85.176,75 = 728,25

▪ JKR = b [$\sum x_i . \sum y_i - \frac{\sum x_i . \sum y_i}{n}$] = (0,897) [(61685)- $\frac{(725).(1011)}{12}$] = [(0,897) [(61.685)- $\frac{(732.975)}{12}$]

- JKR = (0.897) [(61,685) - (61,081.25)] = 7). (603.75) = 541,564
- JKG = JKT-JKR = 728.25 - 541,564 = 186,686
- RJKR =$\frac{JKR}{1} = \frac{541,564}{1} = 541,564$
- RJKG = = = 18,669$\frac{JKG}{n-2}\frac{186,686}{12-2}$

- Test Criteria

  If $|F_{hitung}| \geq F_{tabel}$ then H_o is rejected

  If $|F_{hitung}| < F_{tabel}$ then H_o accepted

- Conclusion

  Because 29.009 > 4.96 → H_0 is rejected

  "H_1" → The independent variable has an effect on the dependent variable

Number of Variations	Sum of Squares	Degrees of Freedom	Average Sum of Squares	F Count
Regression	541,564	1	641,564	29,009
Error	186,686	10	18,669	
Total	728.25	11		

5. <u>Correlation Coefficient between Test Score (X) and Final Score (Y)</u>

To measure the degree of closeness of the relationship, you can use the Pearson correlation formula or the formula below:

$$r_{xy} = \text{b}.\frac{S_x}{S_y} = (0,897).\frac{(7,821)}{(8,137)} = 0,862$$

Based on Guilford's criteria, the relationship between test scores and final scores is **"strong"**

6. **Coefficient of Determination (R^2)**

The coefficient of determination is the square of the correlationr_{xy}

$r_{xy} \rightarrow r_{xy}^2$
r_{xy}^2 = 0,862^2= 0,743 = 74,3%

The test score affects the final score by 74.3%. The rest (100% -74.3% = 25.7%) the final score is influenced by other factors that are not explained in the model

7. Prediction

What is the approximate final grade of a student if the test score is 75?

X = 75

Y = a + bx

Y = 30.053 + 0.897x

Y = 30.053 + 0.897 (75)

Y = 30,053 + 67,275

Y = 97.328

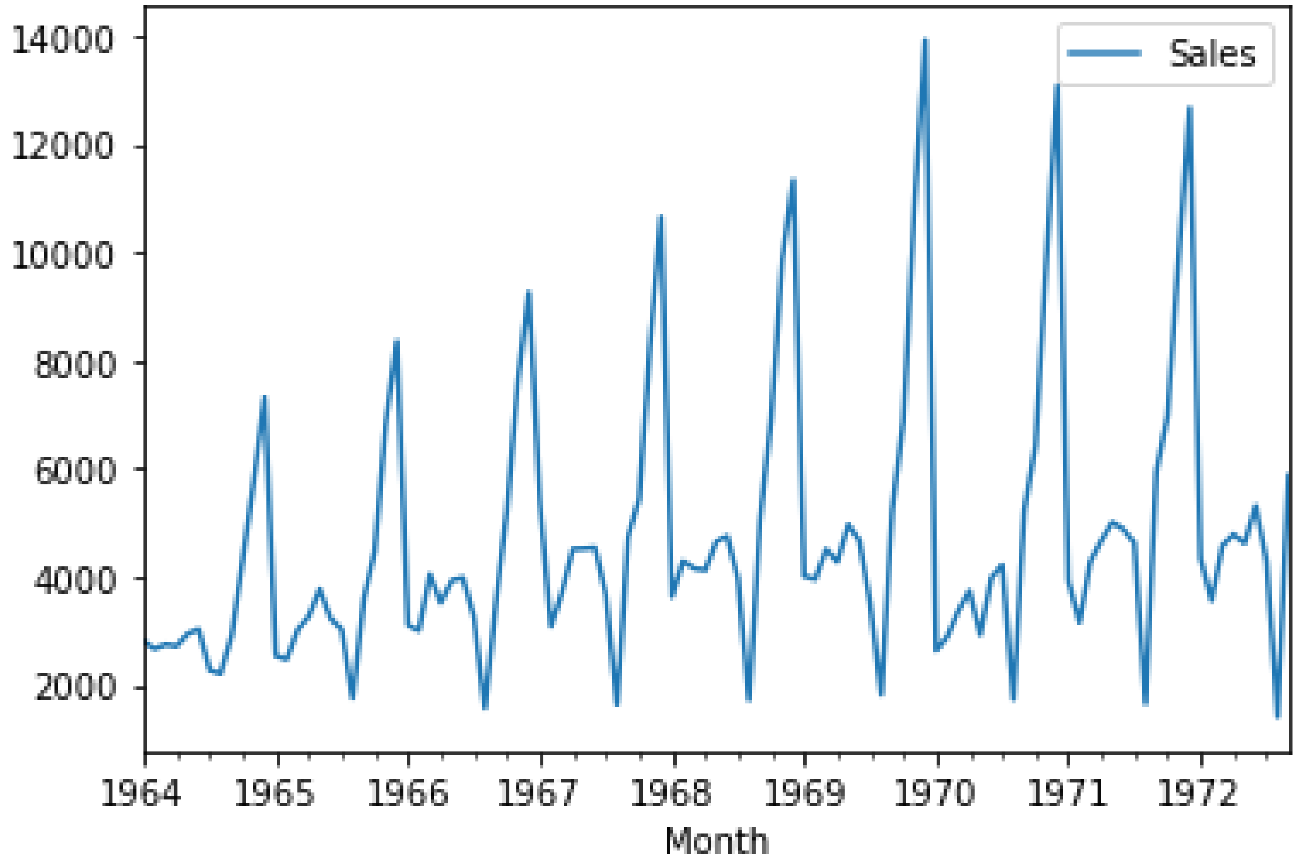
If the student has a test score of 75, the estimated final score is **97.328**

Time Series Model

ARIMA MODEL TO FORECAST FUTURE MONTHLY SALES

The monthly sales data of perrin freres champagne from January 1964 to September 1972 was analyzed for this project. We checked the stationarity of the data , found the autocorrelation factor (***ACF***) and partial autocorrelation factor (***PACF***). We also performed an Augmented ***Dicky-Fuller*** test. We also applied ARIMA Model to make the sales forecast for the upcoming 2 years

Month	Sales	Month	Sales	Month	Sales	Month	Sales	Month	Sales
1964-01	2815	1966-01	3113	1968-01	3633	1970-01	2639	1972-01	4348
1964-02	2672	1966-02	3006	1968-02	4292	1970-02	2899	1972-02	3564
1964-03	2755	1966-03	4047	1968-03	4154	1970-03	3370	1972-03	4577
1964-04	2721	1966-04	3523	1968-04	4121	1970-04	3740	1972-04	4788
1964-05	2946	1966-05	3937	1968-05	4647	1970-05	2927	1972-05	4618
1964-06	3036	1966-06	3986	1968-06	4753	1970-06	3986	1972-06	5312
1964-07	2282	1966-07	3260	1968-07	3965	1970-07	4217	1972-07	4298
1964-08	2212	1966-08	1573	1968-08	1723	1970-08	1738	1972-08	1413
1964-09	2922	1966-09	3528	1968-09	5048	1970-09	5221	1972-09	5877
1964-10	4301	1966-10	5211	1968-10	6922	1970-10	6424		
1964-11	5764	1966-11	7614	1968-11	9858	1970-11	9842		
1964-12	7312	1966-12	9254	1968-12	11331	1970-12	13076		
1965-01	2541	1967-01	5375	1969-01	4016	1971-01	3934		
1965-02	2475	1967-02	3088	1969-02	3957	1971-02	3162		
1965-03	3031	1967-03	3718	1969-03	4510	1971-03	4286		
1965-04	3266	1967-04	4514	1969-04	4276	1971-04	4676		
1965-05	3776	1967-05	4520	1969-05	4968	1971-05	5010		
1965-06	3230	1967-06	4539	1969-06	4677	1971-06	4874		
1965-07	3028	1967-07	3663	1969-07	3523	1971-07	4633		
1965-08	1759	1967-08	1643	1969-08	1821	1971-08	1659		
1965-09	3595	1967-09	4739	1969-09	5222	1971-09	5951		
1965-10	4474	1967-10	5428	1969-10	6872	1971-10	6981		
1965-11	6838	1967-11	8314	1969-11	10803	1971-11	9851		
1965-12	8357	1967-12	10651	1969-12	13916	1971-12	12670		



The monthly data seems to be non stationary and has ***seasonality***

Augmented Dicky-Fuller Test and Differencing

Original Data

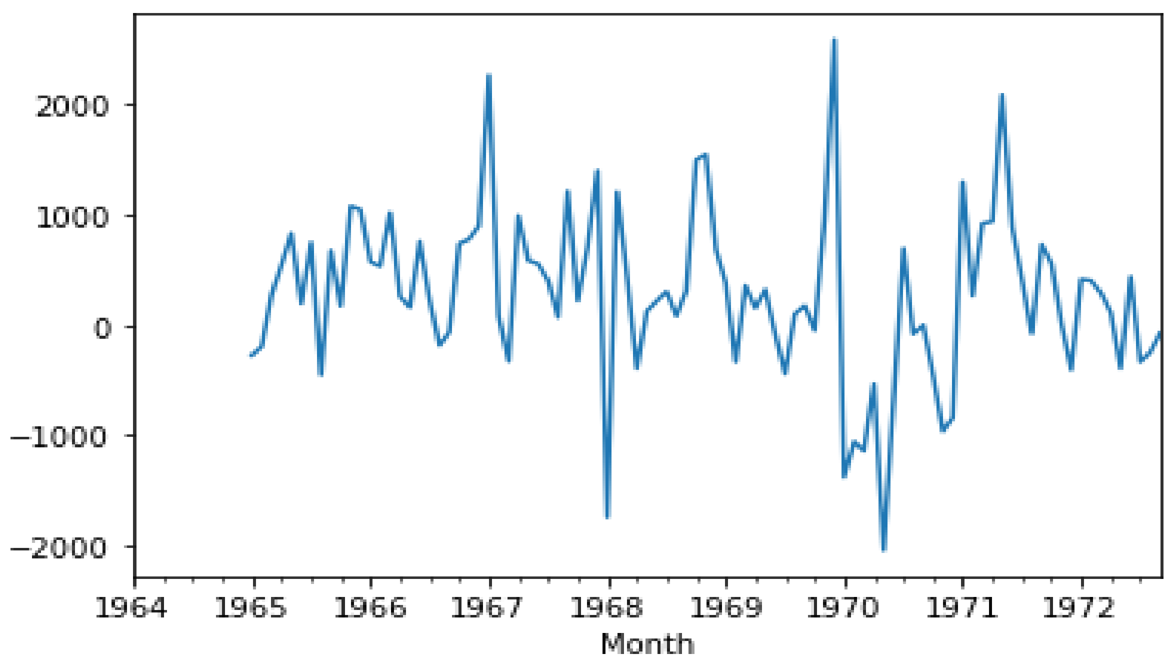
ADF Test Statistic	-1.8335930563276297
p-value	0.3639157716602417
Lags Used	11
Number of Observations Used	93

Since the p value is greater than 0.05 , it indicates that the time series is non stationary and has a unit root. To make the time series data stationary, we used differencing. The plot of sales suggests the presence of seasonality, so we used the seasonal differencing (12 months).

First Seasonal Difference Data

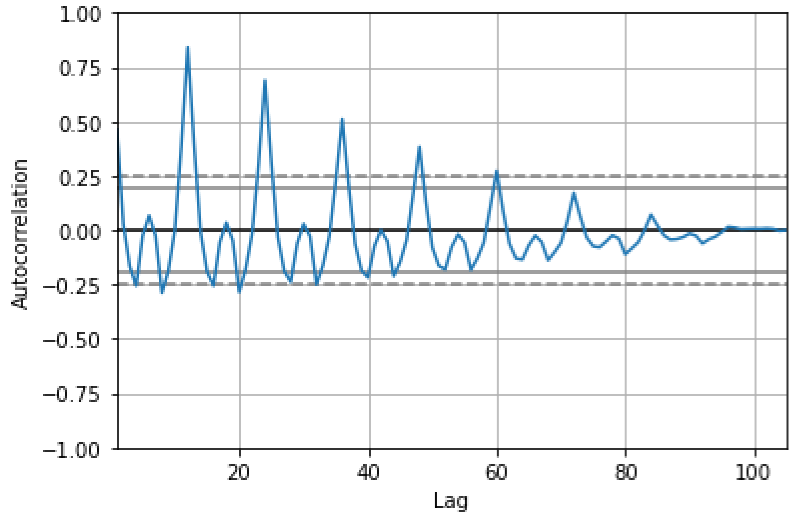
ADF Test Statistic	-7.626619157213163
p-value	2.060579696813685e-11
Lags Used	0
Number of Observations Used	92

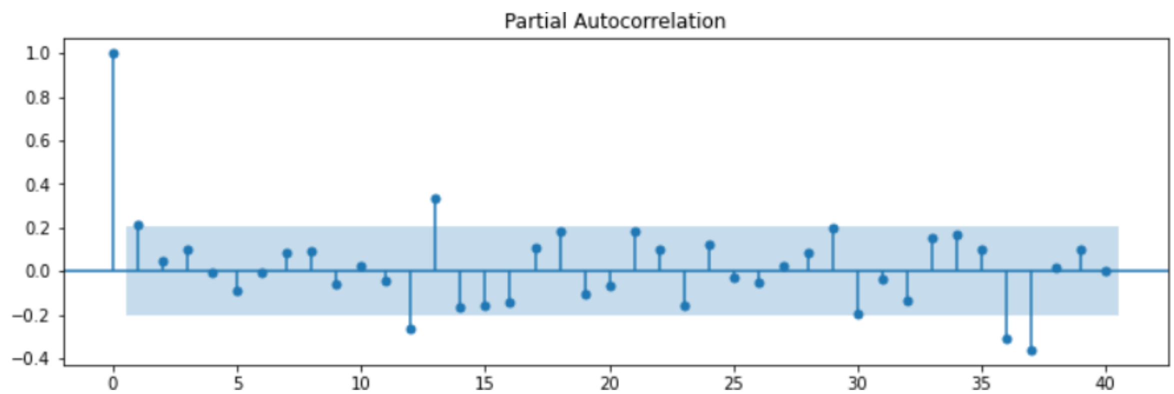
ADF test suggests that the first seasonal difference data is stationary which is also event from the below plot.

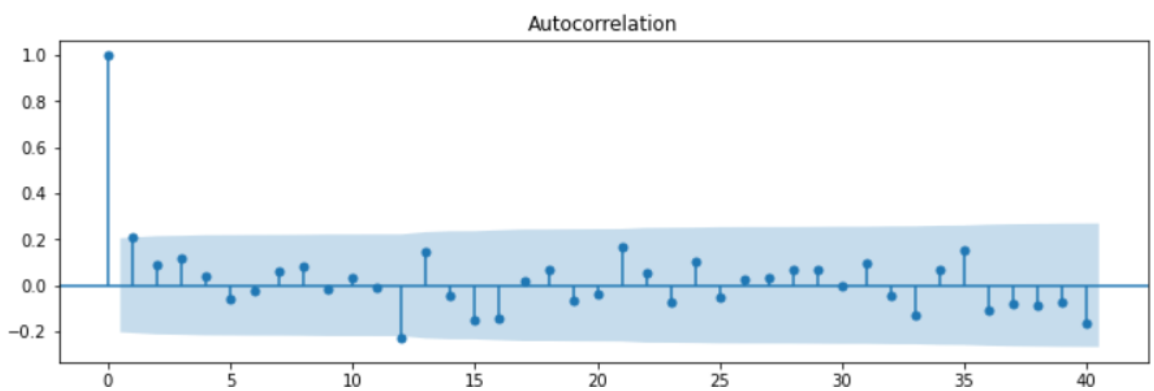


Selecting ARIMA Model Parameters

- d=1 as we did only one seasonal differencing to make the data stationary
- q=1 from the Autocorrelation plot
- p=1 from the Partial Autocorrelation plot







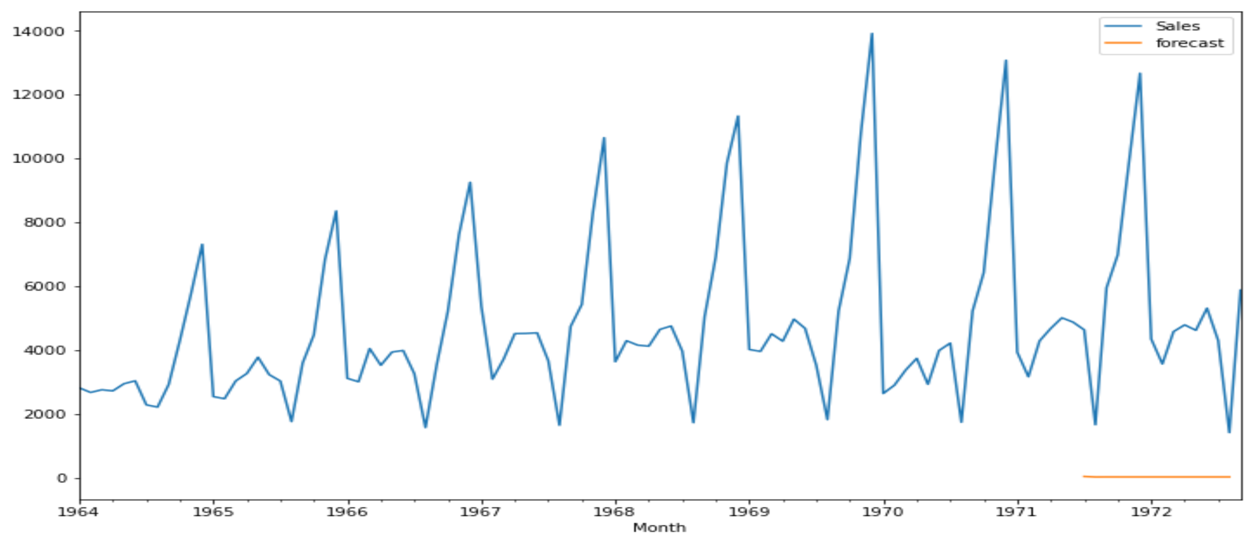
ARIMA Model Results

Dep. Variable:	D.Sales	**No. Observations:**	104
Model:	ARIMA(1, 1, 0)	**Log Likelihood**	-966.440
Method:	css-mle	**S.D. of innovations**	2627.307
Sample:	02-01-1964	**HQIC**	1942.094
	- 09-01-1972		

	coef	std err	z	P>|z|	[0.025	0.975]
const	25.8476	236.330	0.109	0.913	-437.350	489.045
ar.L1.D.Sales	-0.0911	0.099	-0.925	0.355	-0.284	0.102

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-10.9755	+0.0000j	10.9755	0.5000



ARIMA Model didn't give good results and also the p values are very high which suggests that the model is not good.

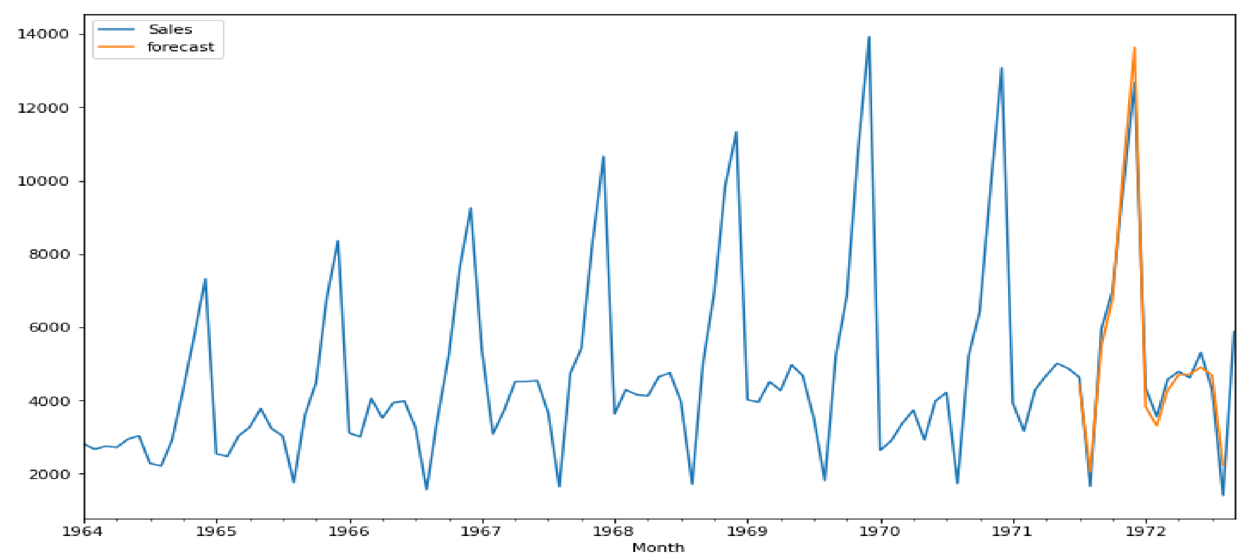
SARIMAX Model Results

SARIMAX Results

Dep. Variable:	Sales	**No. Observations:**	105
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	**Log Likelihood**	-738.402
Date:	Tue, 11 Apr 2023	**AIC**	1486.804
Time:	05:20:13	**BIC**	1499.413
Sample:	01-01-1964	**HQIC**	1491.893
	- 09-01-1972		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	0.2790	0.081	3.433	0.001	0.120	0.438
ma.L1	-0.9494	0.043	-22.334	0.000	-1.033	-0.866
ar.S.L12	-0.4544	0.303	-1.499	0.134	-1.049	0.140
ma.S.L12	0.2450	0.311	0.788	0.431	-0.365	0.855
sigma2	5.055e+05	6.12e+04	8.265	0.000	3.86e+05	6.25e+05

Ljung-Box (L1) (Q):	0.26	**Jarque-Bera (JB):**	8.70
Prob(Q):	0.61	**Prob(JB):**	0.01
Heteroskedasticity (H):	1.18	**Skew:**	-0.21
Prob(H) (two-sided):	0.64	**Kurtosis:**	4.45

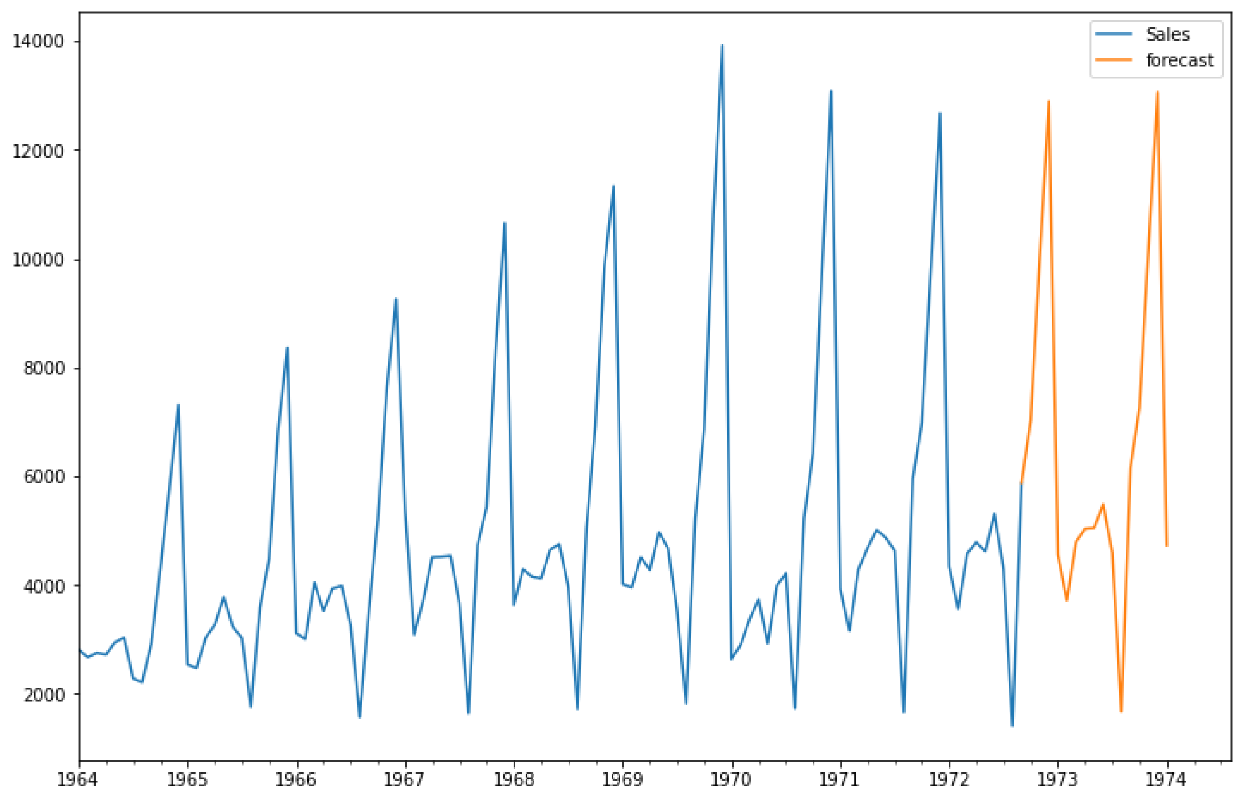


The SARIMAX Model is better than the ARIMA Model which is evident from the comparison of actual sales and the forecasted values . In addition to this, low p values also indicate that the SARIMAX model can be used to forecast the monthly sales .

Forecasting future monthly sales for 2 years

We used the SARIMAX Model to forecast monthly sales for 2 years. Sales forecasting is very important for any company. Some of the benefits of forecasting sales are listed below.

- To predict and plan for demand throughout the year
- To make wise business investments
- To quickly identify And mitigate potential problems
- To improve sales process
- To improve company morale



CODE

In [21]:

```
import pandas as pd
import numpy as np
import statsmodels as sm
from matplotlib import pyplot as plt
%matplotlib inline
```

In [2]:

```
df = pd.read_csv(r"rtsm project.csv")
```

In [3]:

```
df.head()
```

Out[3]:

	Month	Sales
0	1964-01	2815
1	1964-02	2672
2	1964-03	2755
3	1964-04	2721
4	1964-05	2946

In [4]:

```
df.tail()
```

Out[4]:

	Month	Sales
100	1972-05	4618
101	1972-06	5312
102	1972-07	4298
103	1972-08	1413
104	1972-09	5877

In [5]:

```
# Convert Month into Datetime
df['Month']=pd.to_datetime(df['Month'])
```

In [6]:

```
df.set_index('Month',inplace=True)
```

In [7]:

```
df.describe()
```

Out[7]:

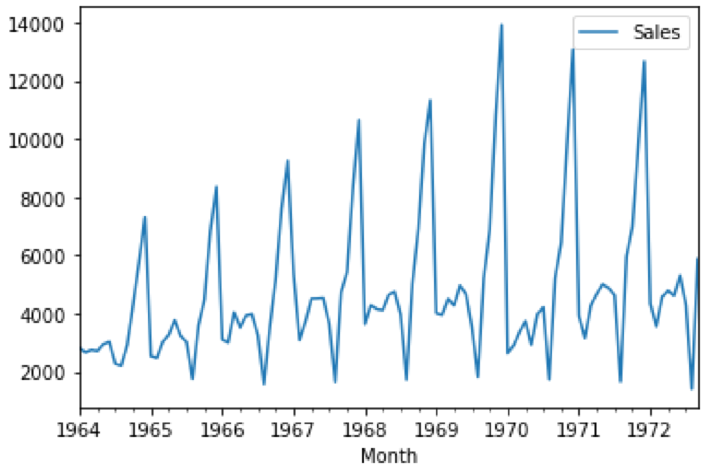
	Sales
count	105.000000
mean	4761.152381
std	2553.502601
min	1413.000000
25%	3113.000000
50%	4217.000000
75%	5221.000000
max	13916.000000

In [8]:

```
df.plot()
```

Out[8]:

```
<AxesSubplot:xlabel='Month'>
```



Testing Stationarity

In [9]:

```
from statsmodels.tsa.stattools import adfuller
test_result=adfuller(df['Sales'])
```

```
C:\Users\abhin\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:7: FutureWar
ning: pandas.Int64Index is deprecated and will be removed from pandas in a future version
. Use pandas.Index with the appropriate dtype instead.
  from pandas import (to_datetime, Int64Index, DatetimeIndex, Period,
C:\Users\abhin\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:7: FutureWar
ning: pandas.Float64Index is deprecated and will be removed from pandas in a future versi
on. Use pandas.Index with the appropriate dtype instead.
  from pandas import (to_datetime, Int64Index, DatetimeIndex, Period,
```

In [10]:

```
#Ho: It is non stationary
#H1: It is stationary

def adfuller_test(sales):
    result=adfuller(sales)
    labels = ['ADF Test Statistic','p-value','#Lags Used','Number of Observations Used']
    for value,label in zip(result,labels):
        print(label+' : '+str(value) )
    if result[1] <= 0.05:
        print("strong evidence against the null hypothesis(Ho), reject the null hypothesi
s. Data has no unit root and is stationary")
    else:
        print("weak evidence against null hypothesis, time series has a unit root, indic
```

```
ating it is non-stationary ")
```

In [11]:

```
adfuller_test(df['Sales'])
```

```
ADF Test Statistic : -1.8335930563276297
p-value : 0.3639157716602417
#Lags Used : 11
Number of Observations Used : 93
weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary
```

Differencing

In [12]:

```
df['Sales First Difference'] = df['Sales'] - df['Sales'].shift(1)
```

In [13]:

```
df['Seasonal First Difference']=df['Sales']-df['Sales'].shift(12)
```

In [14]:

```
df.head(20)
```

Out[14]:

Month	Sales	Sales First Difference	Seasonal First Difference
1964-01-01	2815	NaN	NaN
1964-02-01	2672	-143.0	NaN
1964-03-01	2755	83.0	NaN
1964-04-01	2721	-34.0	NaN
1964-05-01	2946	225.0	NaN
1964-06-01	3036	90.0	NaN
1964-07-01	2282	-754.0	NaN
1964-08-01	2212	-70.0	NaN
1964-09-01	2922	710.0	NaN
1964-10-01	4301	1379.0	NaN
1964-11-01	5764	1463.0	NaN
1964-12-01	7312	1548.0	NaN
1965-01-01	2541	-4771.0	-274.0
1965-02-01	2475	-66.0	-197.0
1965-03-01	3031	556.0	276.0
1965-04-01	3266	235.0	545.0
1965-05-01	3776	510.0	830.0
1965-06-01	3230	-546.0	194.0
1965-07-01	3028	-202.0	746.0
1965-08-01	1759	-1269.0	-453.0

In [15]:

```
## Again test dickey fuller test
adfuller_test(df['Seasonal First Difference'].dropna())
```

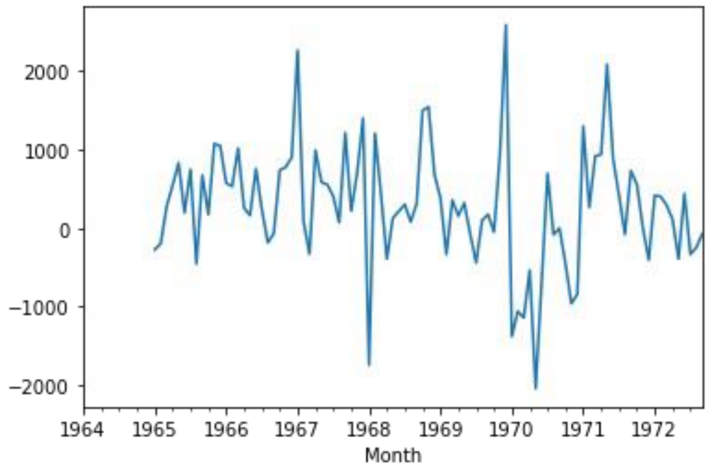
```
ADF Test Statistic : -7.626619157213163
p-value : 2.060579696813685e-11
#Lags Used : 0
Number of Observations Used : 92
strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data has no unit root and is stationary
```

In [16]:

```
df['Seasonal First Difference'].plot()
```

Out[16]:

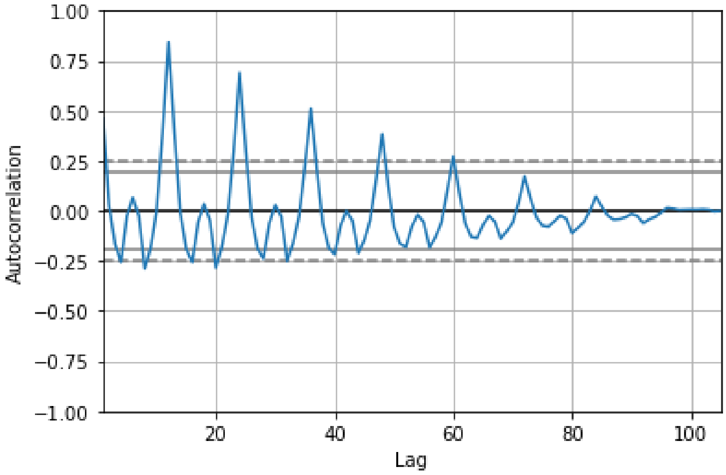
```
<AxesSubplot:xlabel='Month'>
```



AUTO REGRESSIVE MODEL

In [17]:

```
from pandas.plotting import autocorrelation_plot
autocorrelation_plot(df['Sales'])
plt.show()
```

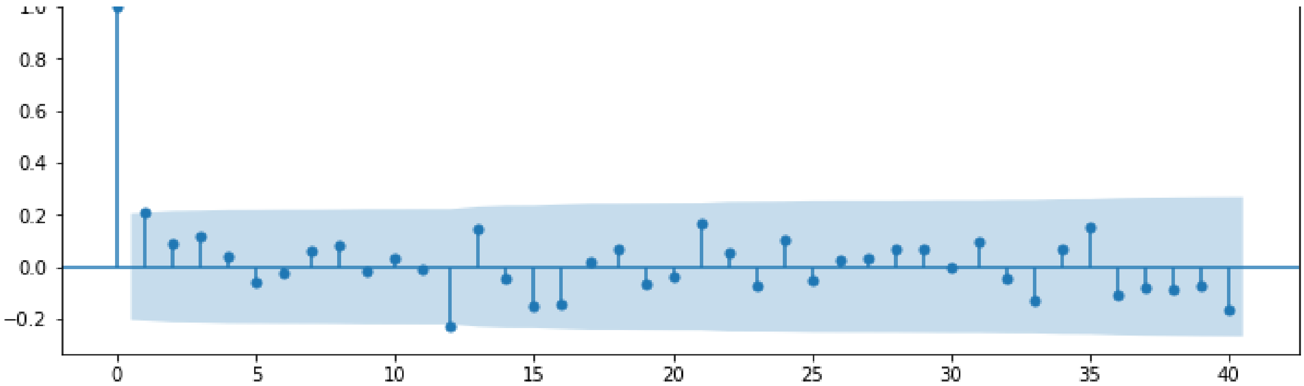


In [18]:

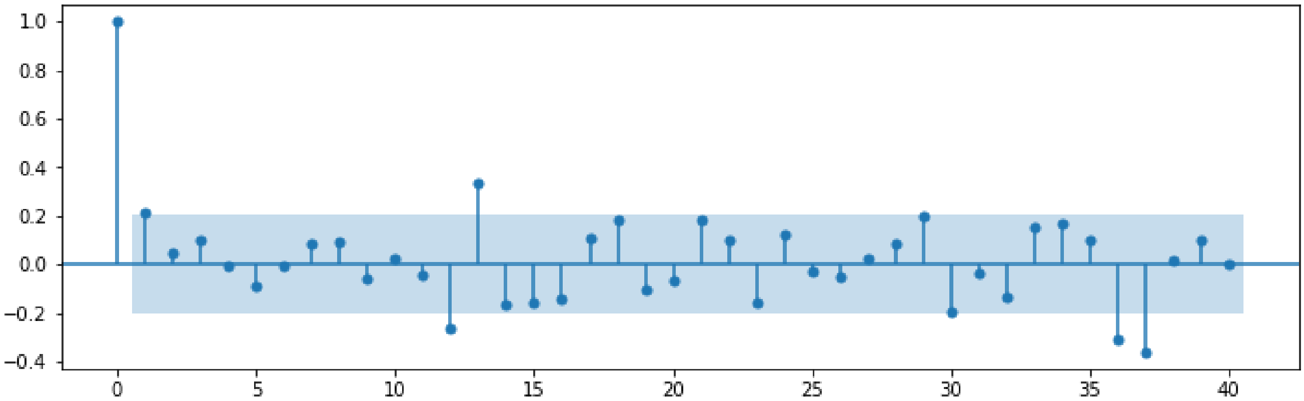
```
from statsmodels.graphics.tsaplots import plot_acf,plot_pacf
```

In [23]:

```
fig = plt.figure(figsize=(12,8))
ax1 = fig.add_subplot(211)
fig = plot_acf(df['Seasonal First Difference'].iloc[13:],lags=40,ax=ax1)
ax2 = fig.add_subplot(212)
fig = plot_pacf(df['Seasonal First Difference'].iloc[13:],lags=40,ax=ax2)
```



Partial Autocorrelation



In [40]:

```
# For non-seasonal data
#p=1, d=1, q=1
from statsmodels.tsa.arima_model import ARIMA
```

In [29]:

```
model=ARIMA(df['Sales'],order=(1,1,1))
```

```
C:\Users\abhin\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:524: ValueWa
rning: No frequency information was provided, so inferred frequency MS will be used.
  warnings.warn('No frequency information was'
C:\Users\abhin\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:524: ValueWa
rning: No frequency information was provided, so inferred frequency MS will be used.
  warnings.warn('No frequency information was'
```

In [30]:

```
model_fit=model.fit()
```

In [31]:

```
model_fit.summary()
```

Out[31]:

ARIMA Model Results

Dep. Variable:	D.Sales	**No. Observations:**	104
Model:	ARIMA(1, 1, 0)	**Log Likelihood**	-966.440
Method:	css-mle	**S.D. of innovations**	2627.307
Date:	Tue, 11 Apr 2023	**AIC**	1938.880
Time:	05:19:06	**BIC**	1946.813
Sample:	02-01-1964	**HQIC**	1942.094
	- 09-01-1972		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	25.8476	236.330	0.109	0.913	-437.350	489.045

ar.L1.D.Sales	-0.0911	0.099	-0.925	0.355	-0.284	0.102

Roots

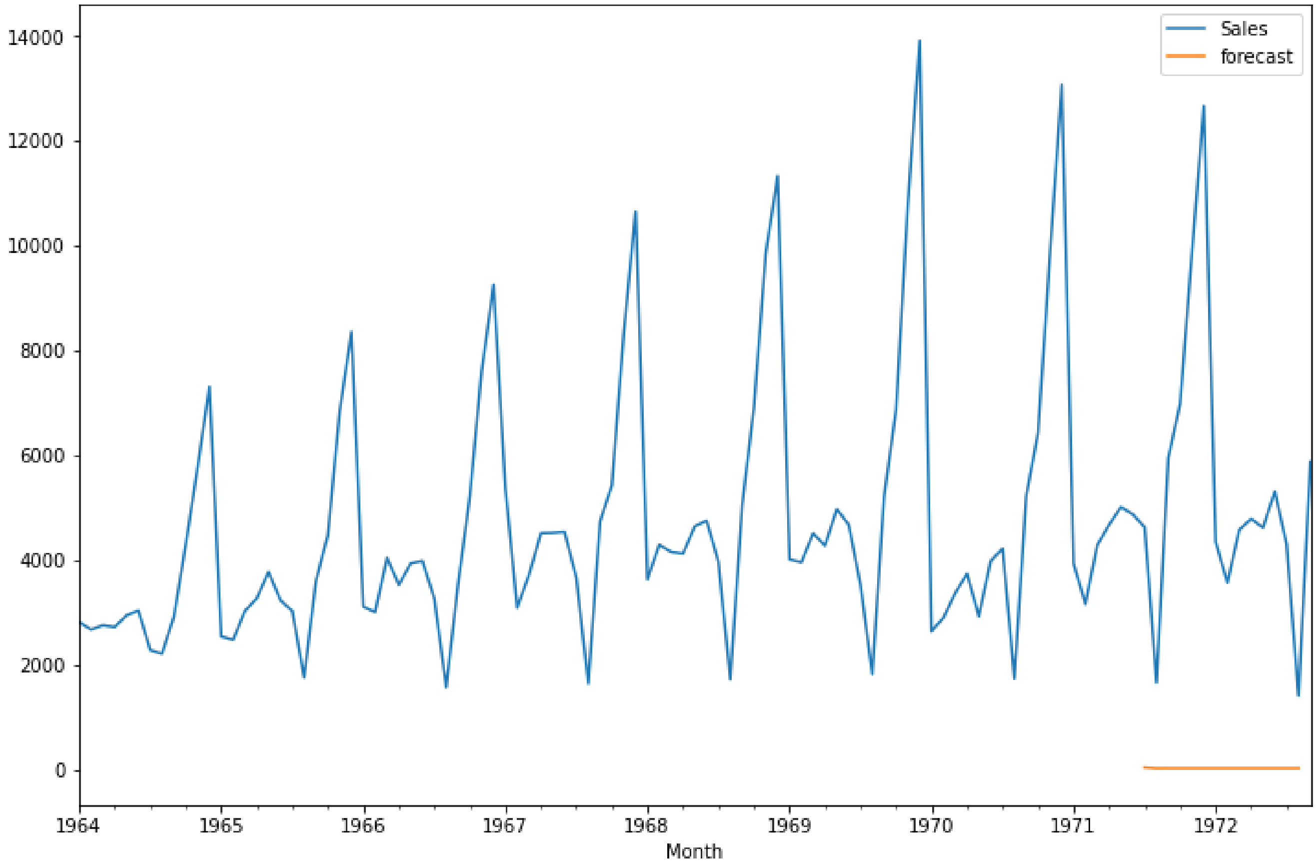
	Real	Imaginary	Modulus	Frequency
AR.1	-10.9755	+0.0000j	10.9755	0.5000

In [32]:

```
df['forecast']=model_fit.predict(start=90,end=103,dynamic=True)
df[['Sales','forecast']].plot(figsize=(12,8))
```

Out[32]:

```
<AxesSubplot:xlabel='Month'>
```



In [41]:

```
import statsmodels.api as sm
model=sm.tsa.statespace.SARIMAX(df['Sales'],order=(1, 1, 1),seasonal_order=(1,1,1,12))
results=model.fit()
```

```
C:\Users\abhin\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:524: ValueWa
rning: No frequency information was provided, so inferred frequency MS will be used.
  warnings.warn('No frequency information was'
C:\Users\abhin\anaconda3\lib\site-packages\statsmodels\tsa\base\tsa_model.py:524: ValueWa
rning: No frequency information was provided, so inferred frequency MS will be used.
  warnings.warn('No frequency information was'
```

In [43]:

```
results.summary()
```

Out[43]:

SARIMAX Results

Dep. Variable:	Sales	**No. Observations:**	105
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 12)	**Log Likelihood**	-738.402

Date:	Tue, 11 Apr 2023	**AIC**	1486.804
Time:	05:19:06	**BIC**	1499.413
Sample:	01-01-1964	**HQIC**	1491.893
	- 09-01-1972		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2790	0.081	3.433	0.001	0.120	0.438
ma.L1	-0.9494	0.043	-22.334	0.000	-1.033	-0.866
ar.S.L12	-0.4544	0.303	-1.499	0.134	-1.049	0.140
ma.S.L12	0.2450	0.311	0.788	0.431	-0.365	0.855
sigma2	5.055e+05	6.12e+04	8.265	0.000	3.86e+05	6.25e+05

Ljung-Box (L1) (Q):	0.26	**Jarque-Bera (JB):**	8.70
Prob(Q):	0.61	**Prob(JB):**	0.01
Heteroskedasticity (H):	1.18	**Skew:**	-0.21
Prob(H) (two-sided):	0.64	**Kurtosis:**	4.45

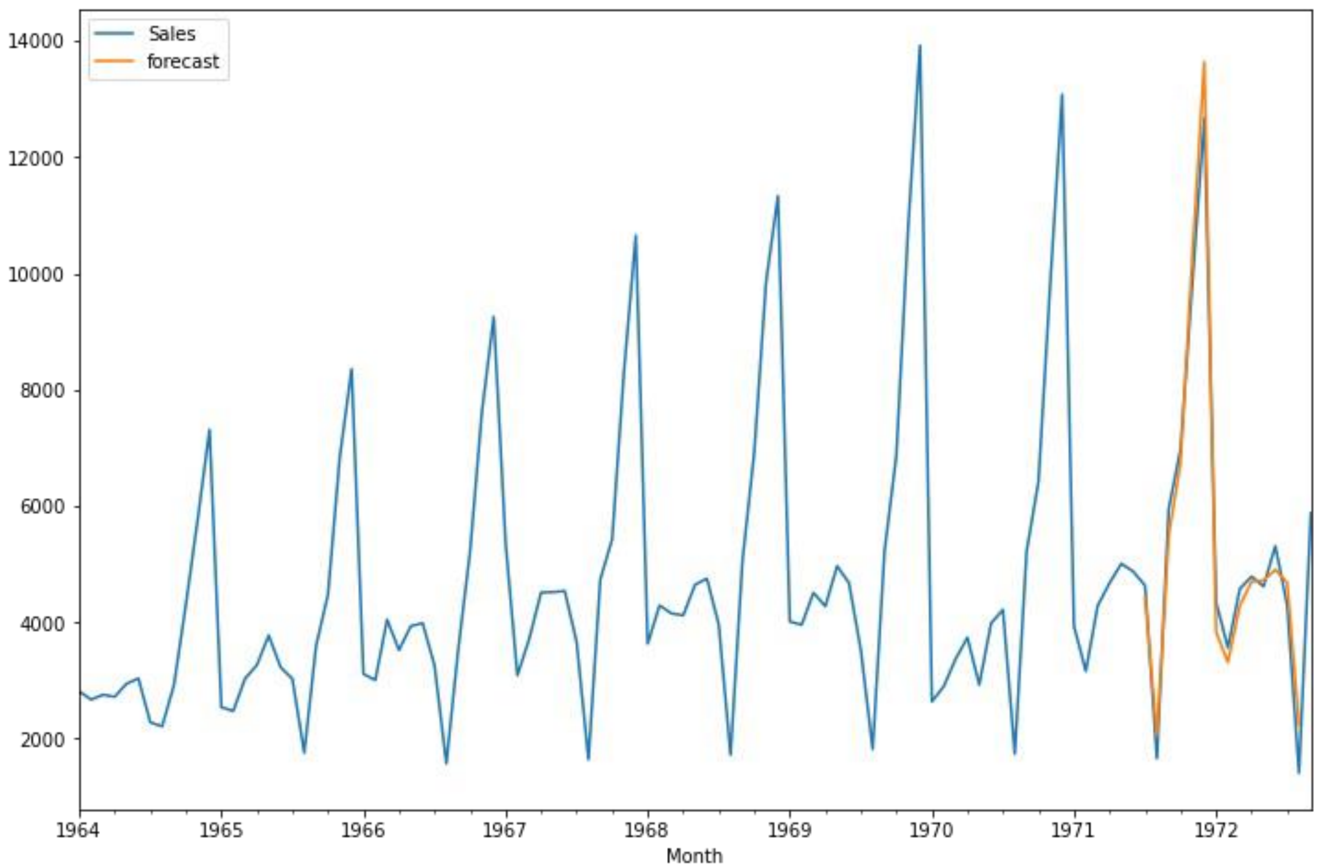
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

In [44]:

```
df['forecast']=results.predict(start=90,end=103,dynamic=True)
df[['Sales','forecast']].plot(figsize=(12,8))
```

Out[44]:

```
<AxesSubplot:xlabel='Month'>
```



In [45]:

```
from pandas.tseries.offsets import DateOffset
```

```
future_dates=[df.index[-1]+ DateOffset(months=x)for x in range(0,24)]
```

In [46]:

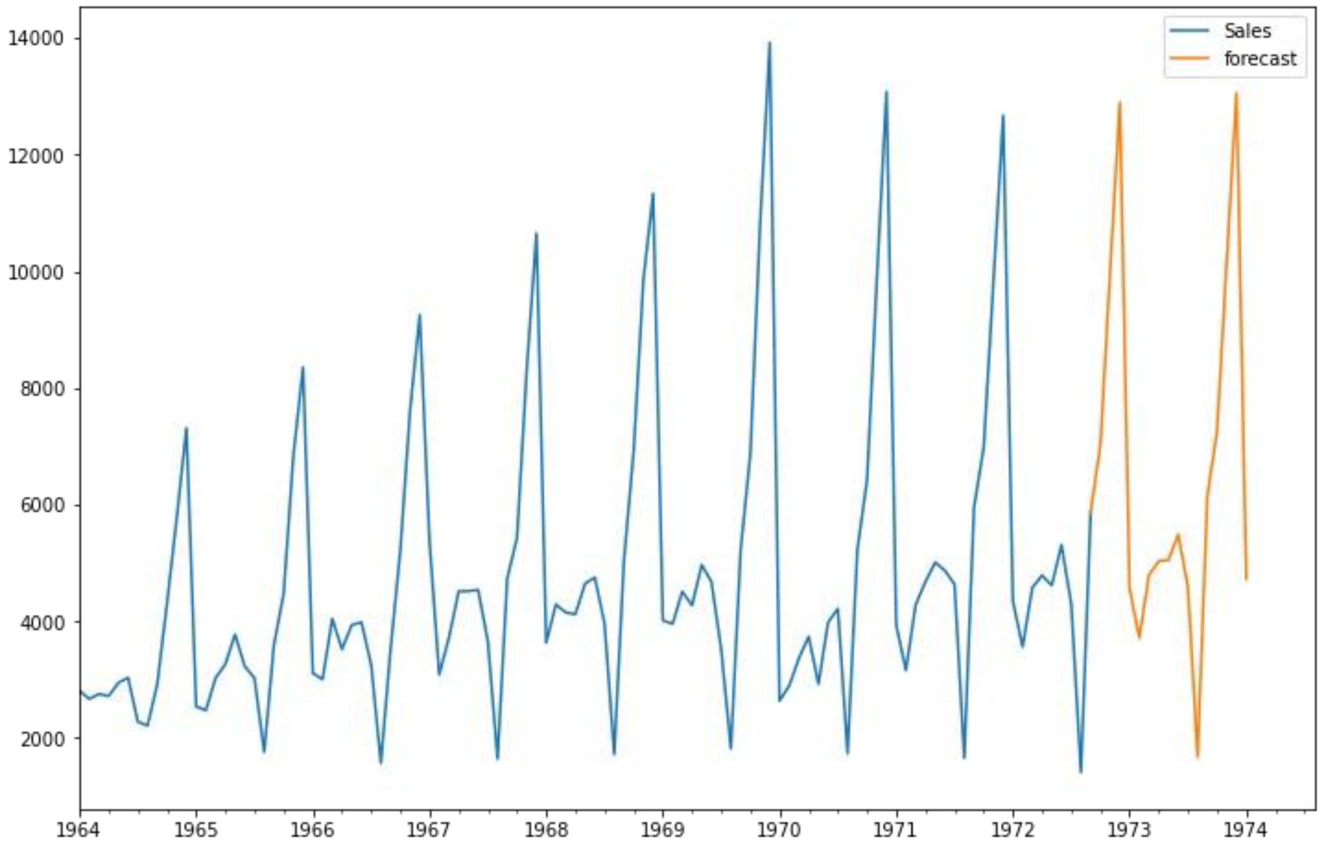
```
future_datest_df=pd.DataFrame(index=future_dates[1:],columns=df.columns)
```

In [47]:

```
future_df=pd.concat([df,future_datest_df])
future_df['forecast'] = results.predict(start = 104, end = 120, dynamic= True)
future_df[['Sales', 'forecast']].plot(figsize=(12, 8))
```

Out[47]:

```
<AxesSubplot:>
```



In [ ]: