1- Data description :

This dataset describes the number of airline passengers per month from 1-1949 to 12-1960 (144 observations)

	#Passengers
count	144.000000
mean	280.298611
std	119.966317
min	104.000000
25%	180.000000
50%	265.500000
75%	360.500000
max	622.000000

	Month	#Passengers
0	1949-01	112
1	1949-02	118
2	1949-03	132
3	1949-04	129
4	1949-05	121
...	...	...
139	1960-08	606
140	1960-09	508
141	1960-10	461
142	1960-11	390
143	1960-12	432

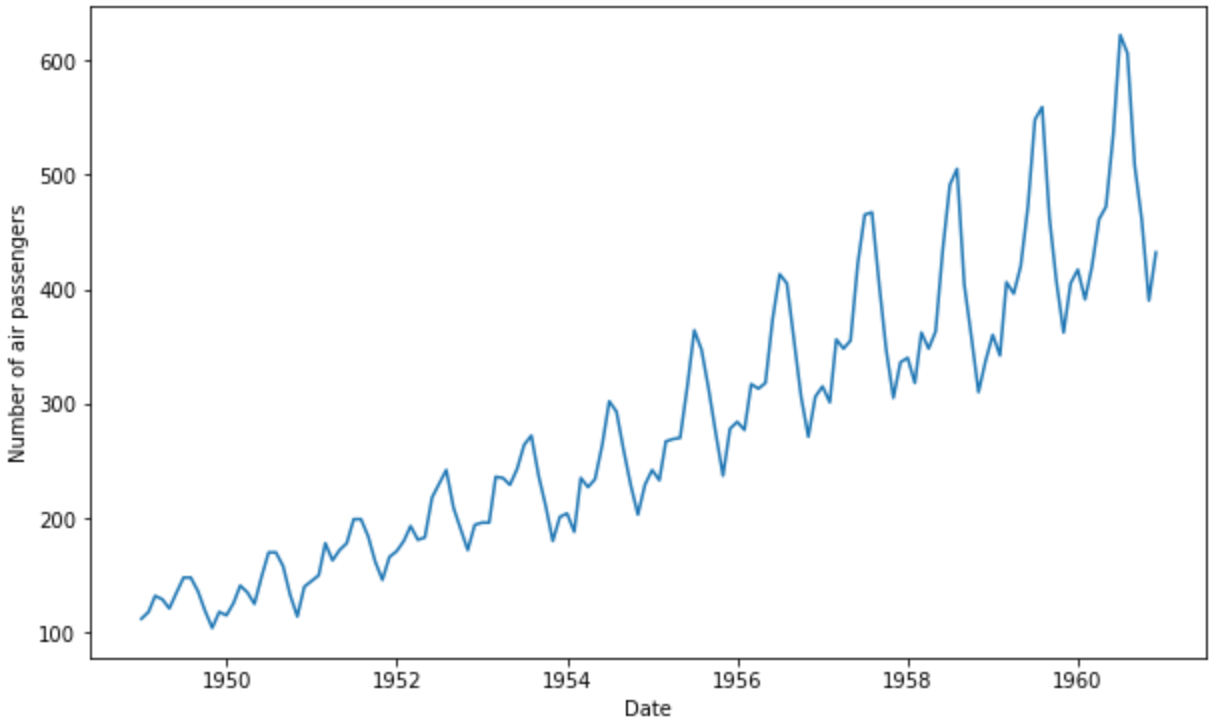
144 rows × 2 columns

2- main objective of this analysis:

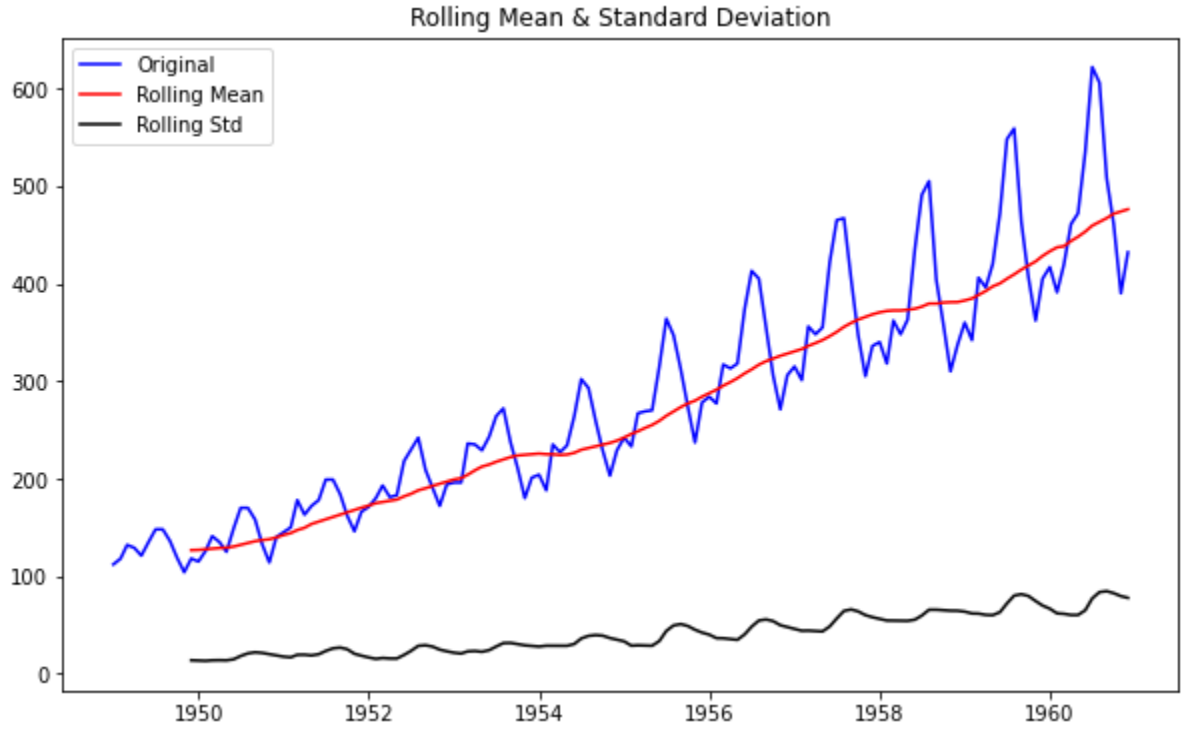
To forecast the number of the passengers for the coming months .

3- EDA:

- Changing the Type of ‘Month’ columns to datetime format from string
- First we made sure we didn’t miss any month in between for continuous observation (144 Months)
- set the index of pandas dataframe to the Month column



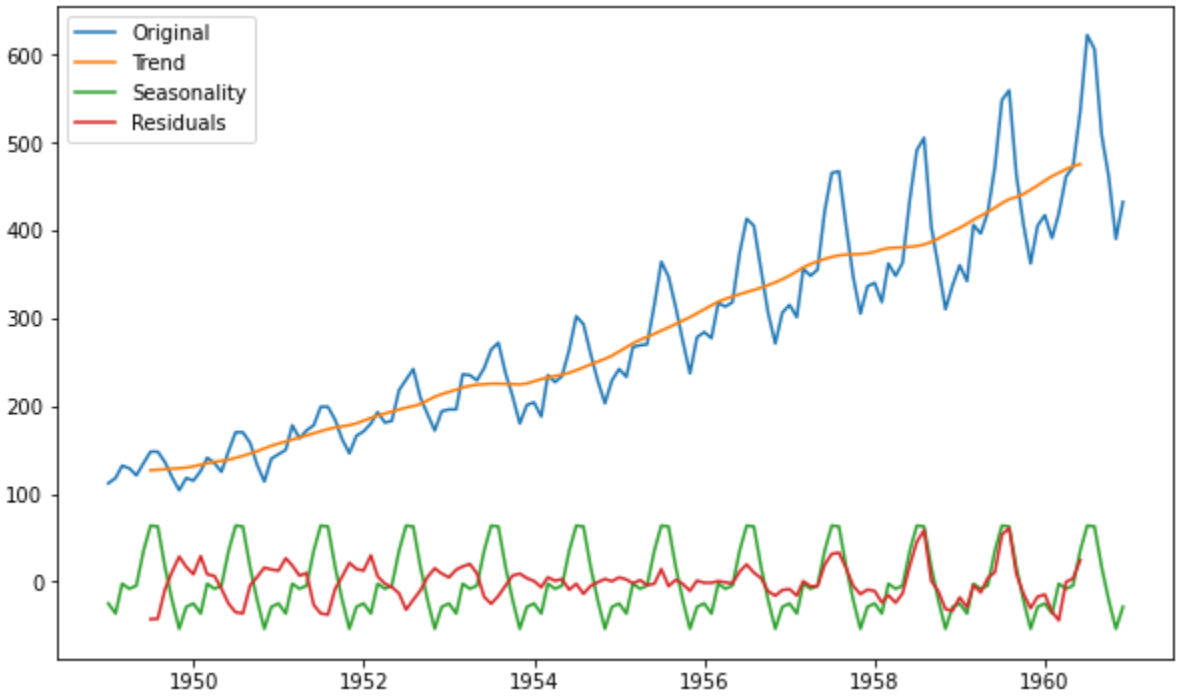
From the graph we notice we have trend and seasonality .



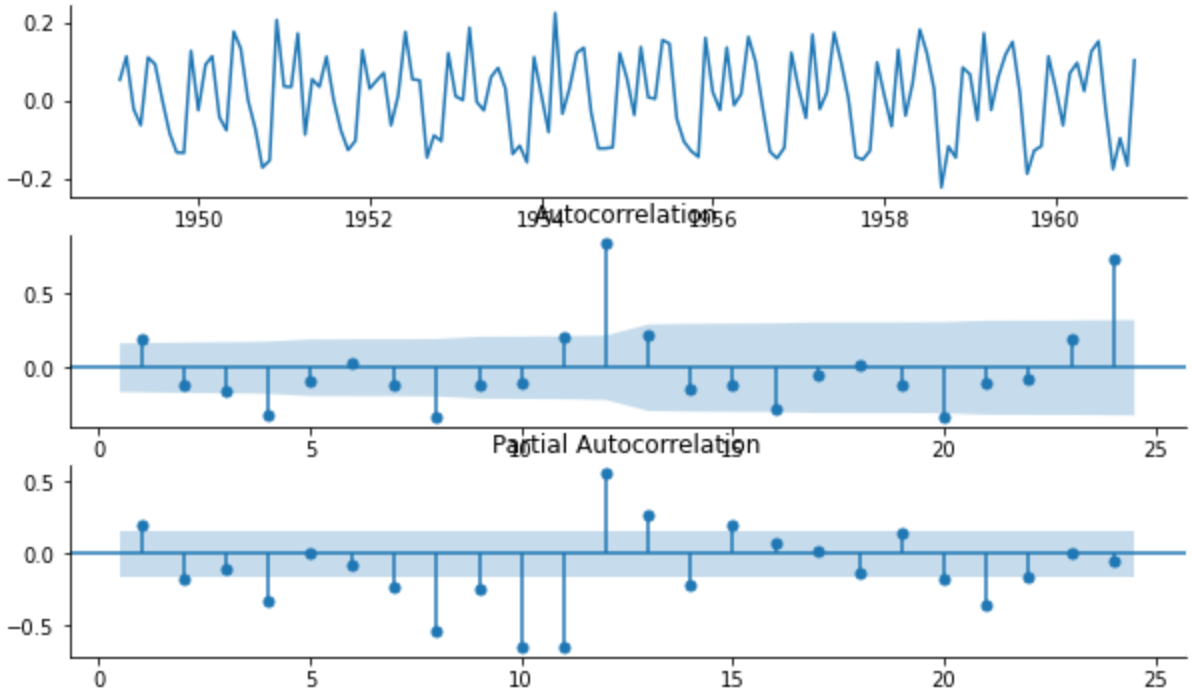
From the graphs shown above the data is not stationary with changing mean and variance. To be more sure we perform the adfuller test :

Results of Dickey Fuller Test:
Test Statistic 0.815369
p-value 0.991880
Lags Used 13.000000
Critical Value (1%) -3.481682
Critical Value (5%) -2.884042
Critical Value (10%) -2.578770
dtype: float64

we can safely say that our Time Series at the moment is not stationary.



To achieve stationarity ,we take the log and shift the data :


0.2
0.0
−0.2
1950
1952
1954
1956
1958
1960
Autocorrelation
0.5
0.0
0
5
10
15
20
25
Partial Autocorrelation
0.5
0.0
−0.5
0
5
10
15
20
25

4- models:

1- SARIMA model with order (1,0,0) and seasonal_order=(0,1,1,12):

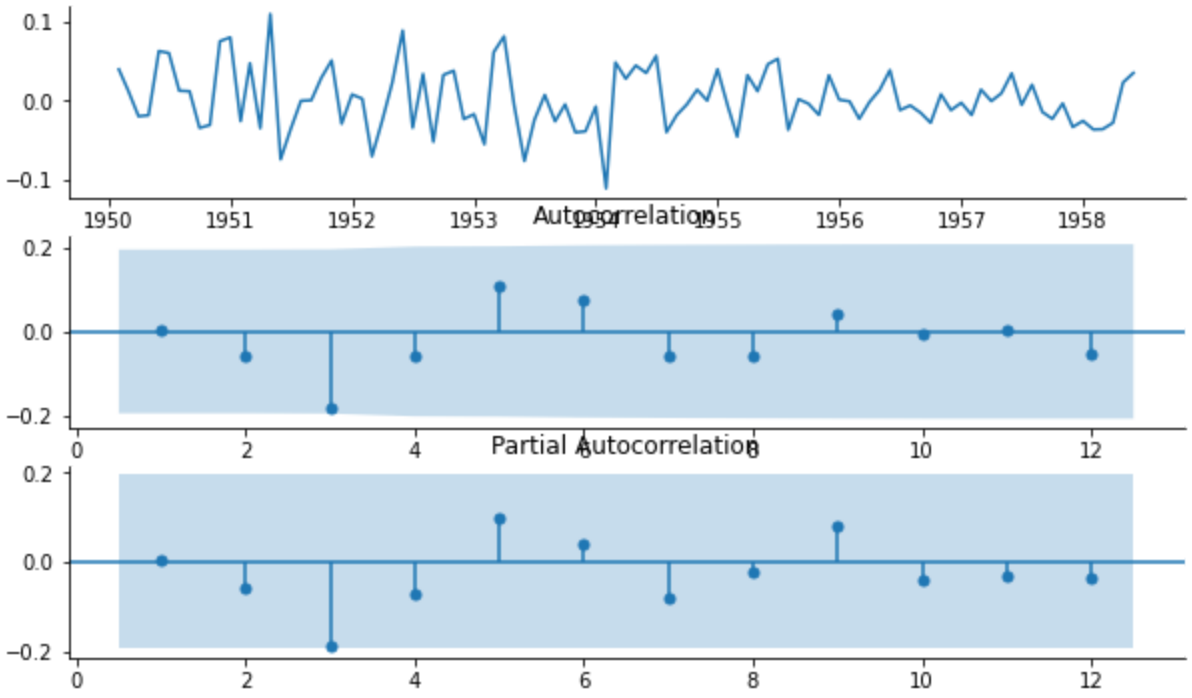
Calculating MAPE = 143.329
Durbin_watson value = 1.99

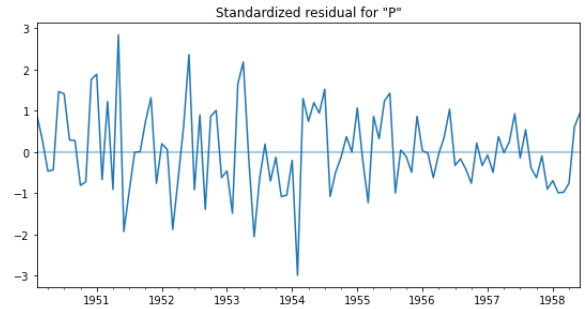
SARIMAX Results

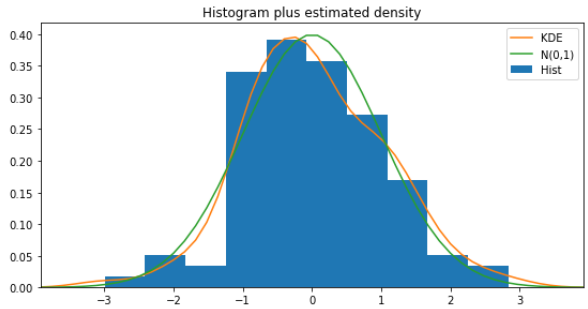
Dep. Variable:	Passengers	**No. Observations:**	113
Model:	SARIMAX(1, 0, 0)x(0, 1, [1], 12)	**Log Likelihood**	186.768
Date:	Mon, 24 May 2021	**AIC**	-365.536
Time:	00:40:55	**BIC**	-355.076
Sample:	02-01-1949	**HQIC**	-361.301
	- 06-01-1958		
Covariance Type:	opg		

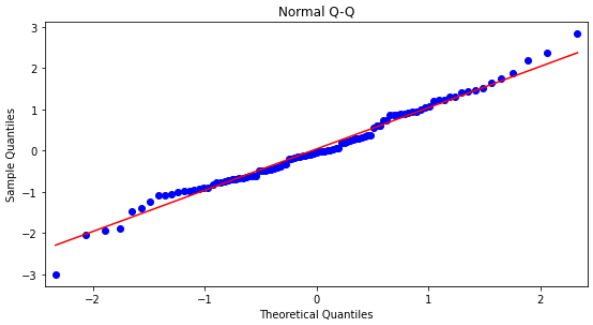
	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.0004	0.002	-0.200	0.842	-0.004	0.004
ar.L1	-0.3008	0.079	-3.802	0.000	-0.456	-0.146
ma.S.L12	-0.5739	0.105	-5.450	0.000	-0.780	-0.368
sigma2	0.0014	0.000	7.345	0.000	0.001	0.002

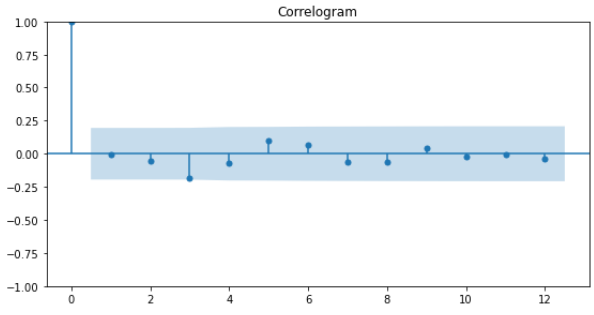
Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	0.65
Prob(Q):	0.98	**Prob(JB):**	0.72
Heteroskedasticity (H):	0.23	**Skew:**	0.12
Prob(H) (two-sided):	0.00	**Kurtosis:**	3.31

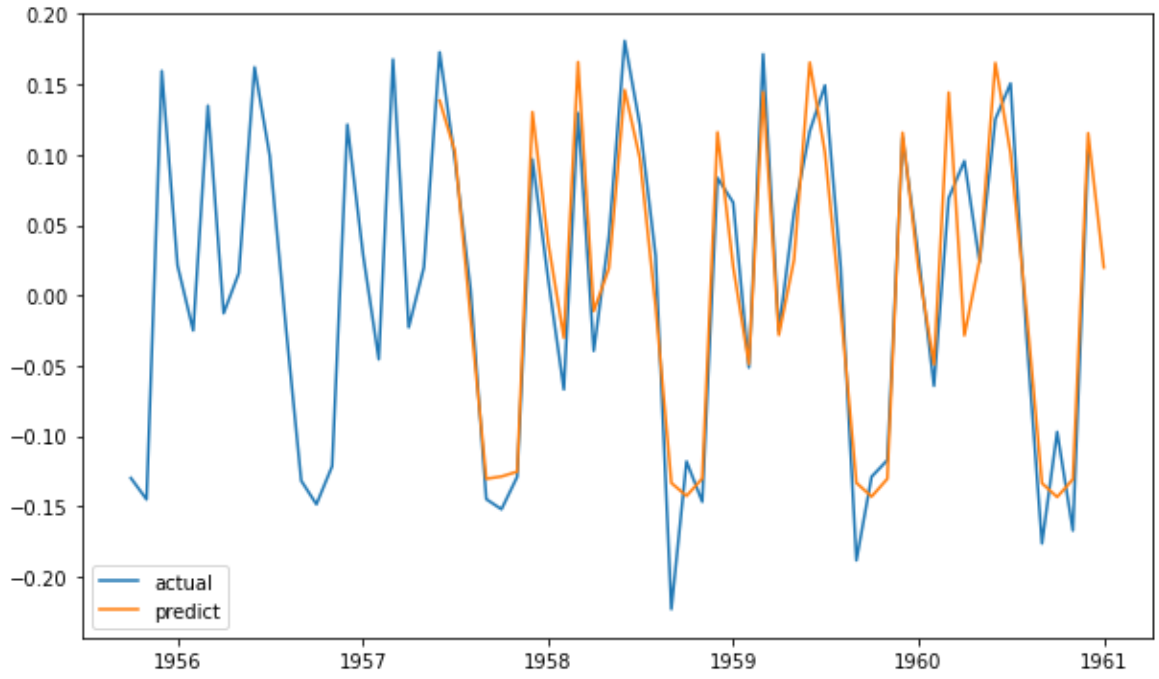

Autocorrelation
Partial Autocorrelation


Standardized residual for "P"


Histogram plus estimated density
KDE
N(0,1)
Hist

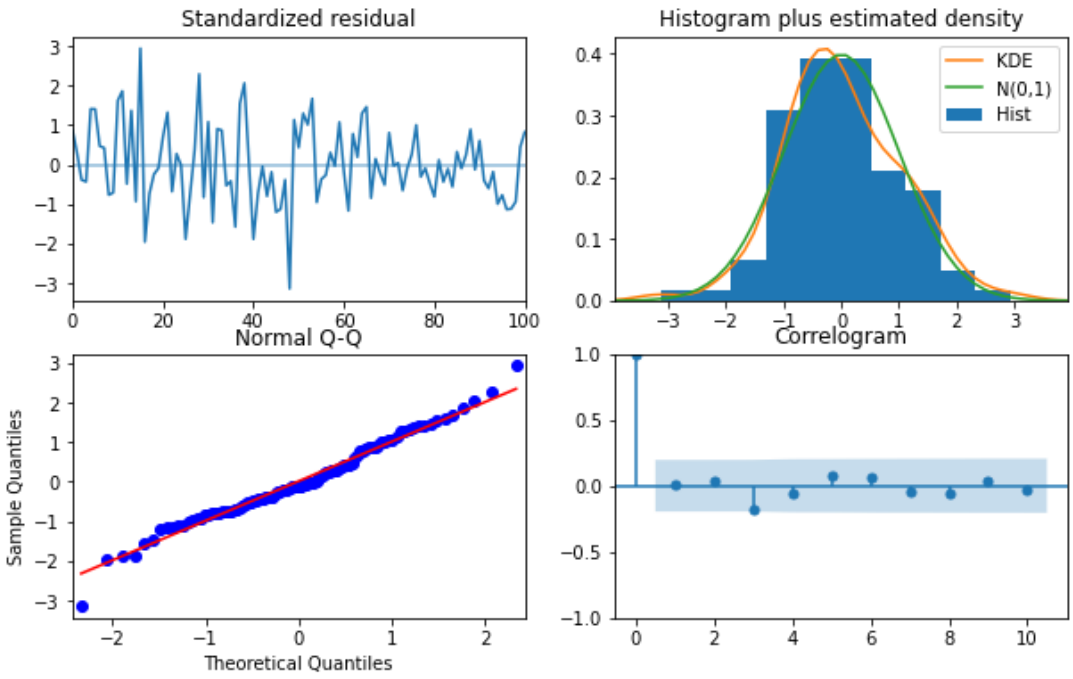

Normal Q-Q
Sample Quantiles
Theoretical Quantiles

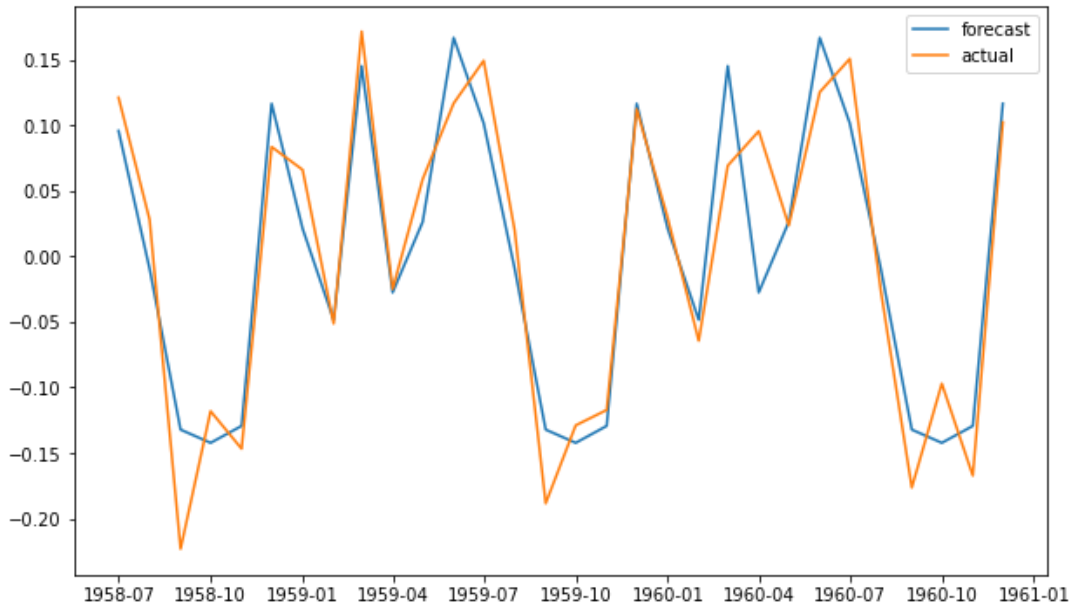

Correlogram



2- using pyramid.arima - auto_arima model with order (1,0,0) and seasonal_order=(0,1,1,12):

Best model: ARIMA(0,0,1)(0,1,1)[12]
Total fit time: 9.488 seconds
-367.88350216500305
MAPE = 148.12





5- key findings:

Although the MAPE value is more for the auto_arima model ,but it seems to be doing a better job predicting closer values to the actual values .

6- Suggestions for next steps:

Providing more data points will help train the models better for forecasting , maybe tying more different values for the p and q values for further investigation and improvement.