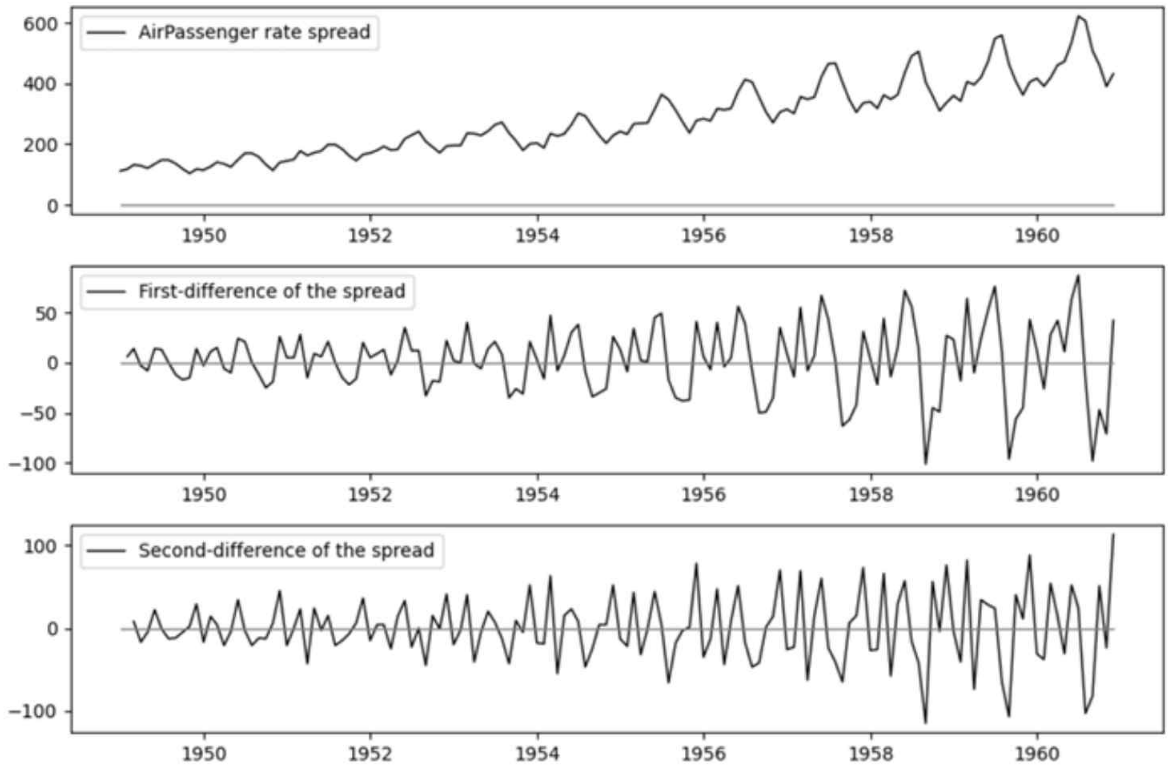
Time Series Analysis and Forecasting

[Individual Project Report]

Industrial Engineering_2020067356_장윤수

● **Please check the stationarity. If needed, you may need to use a d-th difference of yt or ARIMA(p,d,q)**



1) In air passenger rate spread

- There exists a trend. Hence, It is not a covariance stationary.

2) In first-difference air passenger rate spread

- Mean of 0 but It seems to have heteroskedasticity.

3) second-difference air passenger rate spread

- Mean of 0 but It seems to have heteroskedasticity.

ADF Test

```
Statistics: 0.8153688792060498
p-value: 0.991880243437641
Critical values: {'1%': -3.4816817173418295, '5%': -2.8840418343195267, '10%': -2.578770059171598}
-----------------------------------------------------------------------------------------------------------
Statistics: -2.8292668241700047
p-value: 0.05421329028382478
Critical values: {'1%': -3.4816817173418295, '5%': -2.8840418343195267, '10%': -2.578770059171598}
-----------------------------------------------------------------------------------------------------------
Statistics: -16.384231542468513
p-value: 2.7328918500142026e-29
Critical values: {'1%': -3.4816817173418295, '5%': -2.8840418343195267, '10%': -2.578770059171598}
```

KPSS Test

```
Statistics: 1.6513122354165206
bounded p-value: 0.01
Critical values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
-----------------------------------------------------------------------------------------------------------
Statistics: 0.023897614400183967
bounded p-value: 0.1
Critical values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
-----------------------------------------------------------------------------------------------------------
Statistics: -16.384231542468513
p-value: 2.7328918500142026e-29
Critical values: {'1%': -3.4816817173418295, '5%': -2.8840418343195267, '10%': -2.578770059171598}
```

1) In air passenger rate spread

- In ADF test, p-value is almost 1 > 0.05 so that we can not reject null hypothesis.

- In KPSS test, p-value is 0.01<0.05, so that we can reject the null hypothesis.

-> Not covariance stataionary.

2) In first-difference air passenger rate spread

- In ADF test, p-value is 0.054 >= 0.05 so that we may reject null hypothesis.

- In KPSS test, p-value is 0.1 > 0.05, so that we can not reject the null hypothesis.

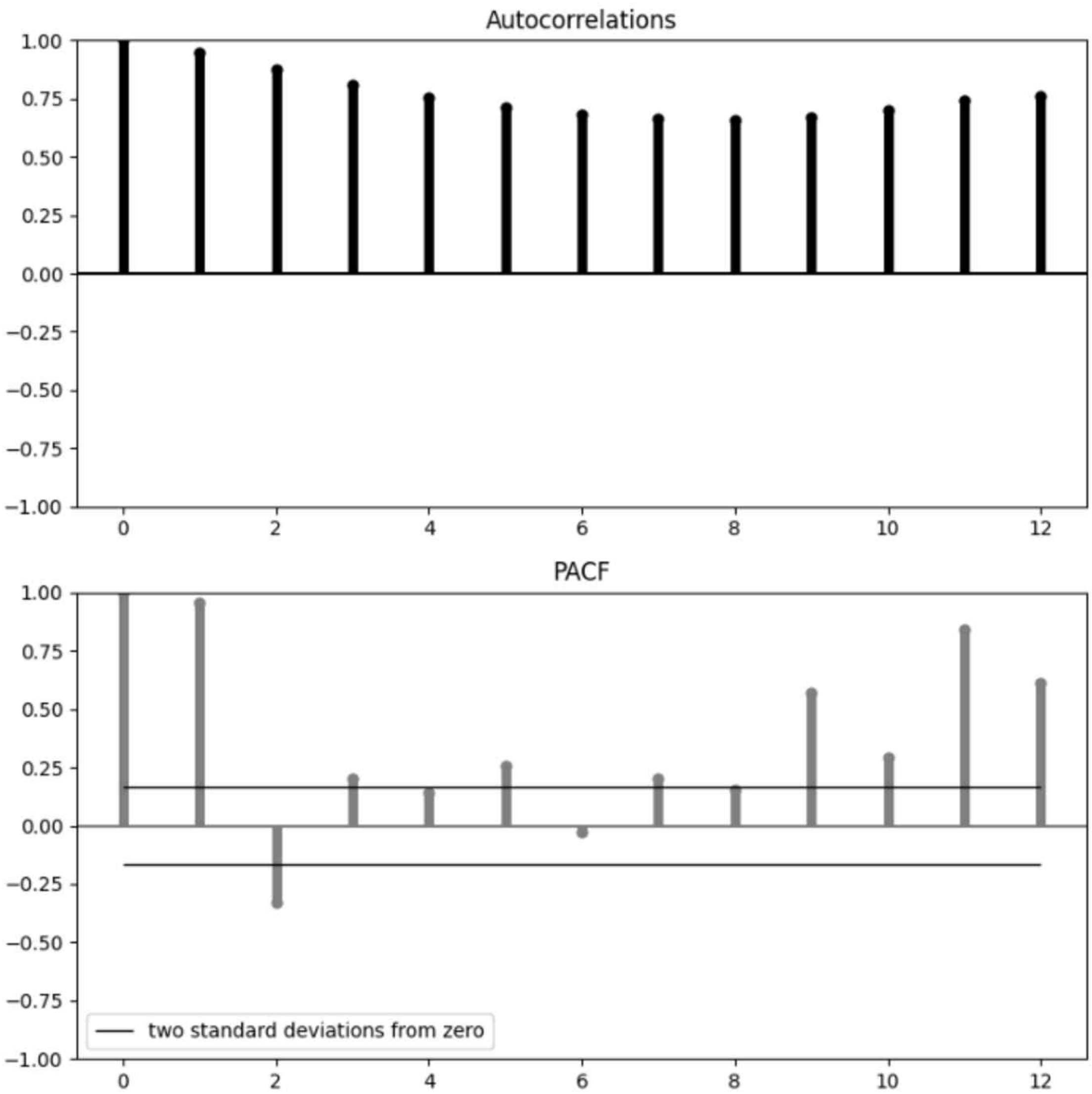
3) second-difference air passenger rate spread

- In ADF test, p-value is 2.73 > 0.05 so that we can not reject null hypothesis.

- In KPSS test, p-value is 2.73 > 0.05, so that we can not reject the null hypothesis.

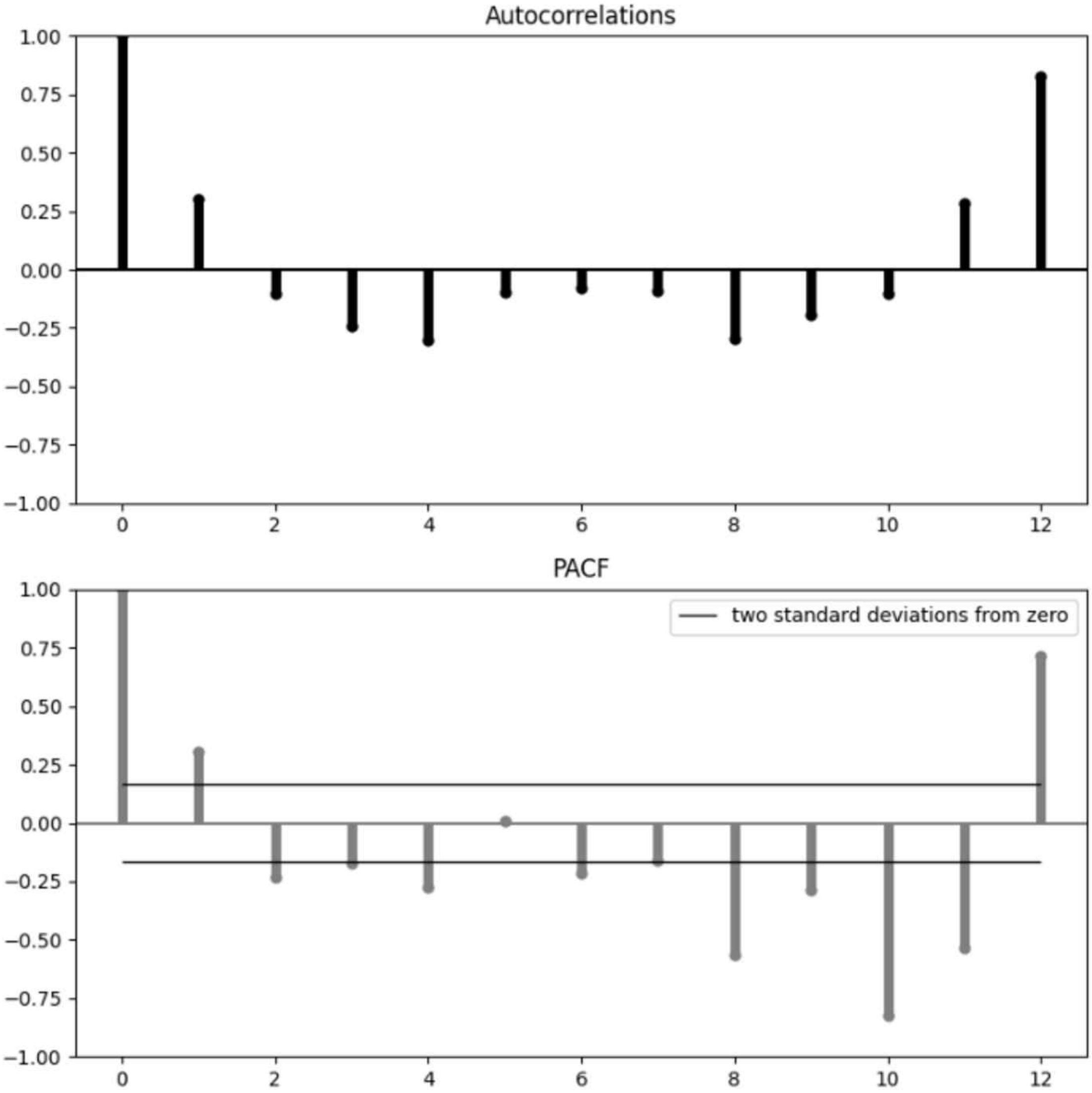
In this point, we can conclude that first-difference air passenger rate spread is covariance stationary and I will use this data. Effect of double-differencing would be too heavy on the data and might result in poor performance.

● **Based on ACF and PACF, suggest at least three candidate models with appropriate description.**



1) ACF & PACF in air passenger rate spread

-> ACF does not converge to 0, it is clear that this data cannot be used.



2) ACF & PACF in first-difference air passenger rate spread

Based on the graph, we can suggest eight candidates.

- ARIMA(1,1,0)　　- ARIMA(2,1,0)
- ARIMA(1,1,1)　　- ARIMA(2,1,1)
- ARIMA(1,1,2)　　- ARIMA(2,1,2)
- ARIMA(1,1,3)　　- ARIMA(2,1,3)

	(1, 1, 3)
ar.L1	{'coef': 1.0, 't_stats': 136.15}
ar.L2	NaN
ma.L1	{'coef': -1.1, 't_stats': -0.56}
ma.L2	{'coef': -0.79, 't_stats': -3.44}
ma.L3	{'coef': 0.89, 't_stats': 0.51}
SSE	150594.2
AIC	1643.36
SBC	1657.96
Q(4)	{'q_stats': 36.55, 'p_val': 0.0}
Q(8)	{'q_stats': 81.32, 'p_val': 0.0}
Q(12)	{'q_stats': 181.51, 'p_val': 0.0}

	(2, 1, 2)	(2, 1, 3)
ar.L1	{'coef': 0.98, 't_stats': 49.17}	{'coef': -0.1, 't_stats': -0.26}
ar.L2	{'coef': -0.98, 't_stats': -53.33}	{'coef': 0.44, 't_stats': 2.91}
ma.L1	{'coef': -0.85, 't_stats': -0.44}	{'coef': 0.38, 't_stats': 0.95}
ma.L2	{'coef': 1.0, 't_stats': 0.22}	{'coef': -0.8, 't_stats': -8.47}
ma.L3	NaN	{'coef': -0.29, 't_stats': -1.11}
SSE	128015.07	141256.03
AIC	1621.11	1636.59
SBC	1635.71	1654.11
Q(4)	{'q_stats': 5.36, 'p_val': 0.25}	{'q_stats': 13.43, 'p_val': 0.01}
Q(8)	{'q_stats': 52.53, 'p_val': 0.0}	{'q_stats': 27.05, 'p_val': 0.0}
Q(12)	{'q_stats': 148.75, 'p_val': 0.0}	{'q_stats': 114.43, 'p_val': 0.0}

(The t-statistics at 5% significance level is 1.93.)

For ARIMA(1,1,3), ARIMA(2,1,2), ARIMA(2,1,3), they can not show statisticaclly significant corfficient for some lags.

	(2, 1, 0)	(1, 1, 0)
ar.L1	{'coef': 0.38, 't_stats': 4.18}	{'coef': 0.31, 't_stats': 3.68}
ar.L2	{'coef': -0.23, 't_stats': -3.34}	NaN
ma.L1	NaN	NaN
ma.L2	NaN	NaN
ma.L3	NaN	NaN
SSE	161219.47	168447.42
AIC	1648.7	1652.71
SBC	1657.46	1658.55
Q(4)	{'q_stats': 13.49, 'p_val': 0.01}	{'q_stats': 16.46, 'p_val': 0.0}
Q(8)	{'q_stats': 30.87, 'p_val': 0.0}	{'q_stats': 26.64, 'p_val': 0.0}
Q(12)	{'q_stats': 125.05, 'p_val': 0.0}	{'q_stats': 130.82, 'p_val': 0.0}

	(1, 1, 1)
ar.L1	{'coef': -0.48, 't_stats': -3.74}
ar.L2	NaN
ma.L1	{'coef': 0.87, 't_stats': 10.84}
ma.L2	NaN
ma.L3	NaN
SSE	158910.81
AIC	1646.73
SBC	1655.49
Q(4)	{'q_stats': 16.3, 'p_val': 0.0}
Q(8)	{'q_stats': 27.49, 'p_val': 0.0}
Q(12)	{'q_stats': 116.84, 'p_val': 0.0}

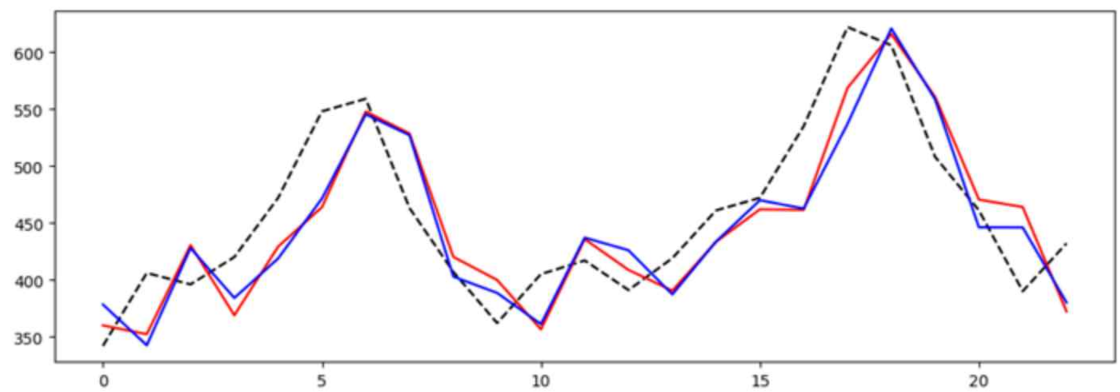
For ARIMA(2,1,0), ARIMA(1,1,0), ARIMA(1,1,1), they can show statisticaclly significant corfficient but have larger AIC and SBC than ARIMA(1,1,2) and ARIMA(2,1,1).

(1, 1, 2)	(2, 1, 1)
{'coef': 0.58, 't_stats': 5.74}	{'coef': 1.09, 't_stats': 12.71}
NaN	{'coef': -0.49, 't_stats': -4.84}
'coef': -0.32, 't_stats': -3.12}	'coef': -0.84, 't_stats': -11.06}
'coef': -0.51, 't_stats': -7.06}	NaN
NaN	NaN
148682.5	142252.98
1639.61	1633.55
1651.29	1645.23
{'q_stats': 14.5, 'p_val': 0.01}	{'q_stats': 9.58, 'p_val': 0.05}
{'q_stats': 34.35, 'p_val': 0.0}	{'q_stats': 21.61, 'p_val': 0.01}
'q_stats': 124.27, 'p_val': 0.0}	{'q_stats': 115.05, 'p_val': 0.0}

At this point, we have two candidates, ARIMA(1,1,2) and ARIMA(2,1,1).

● **Use the first 10 years for estimation and rest of 2 years for one-step ahead forecasting.**

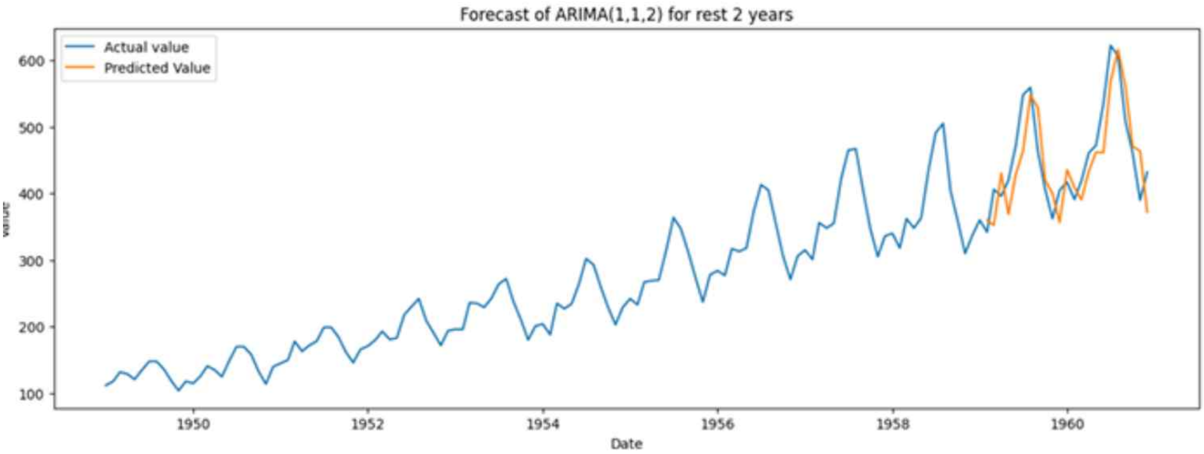
●**[Extra 5 Points for Total Score in Class] Add seasonality into the ARIMA model (Seasonal ARIMA)**



Red: ARIMA(1,1,2)

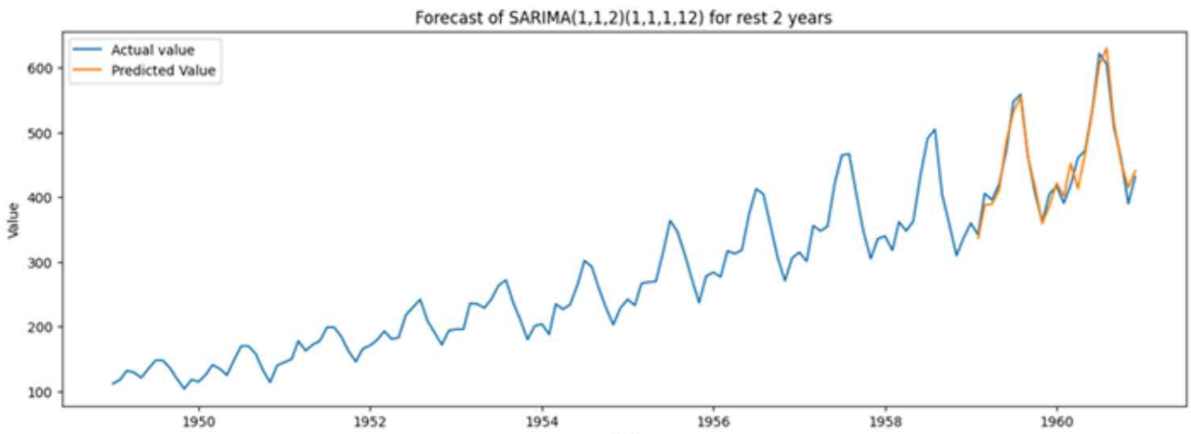
Blue: ARIMA(2,1,1)

1) ARIMA(1,1,2)

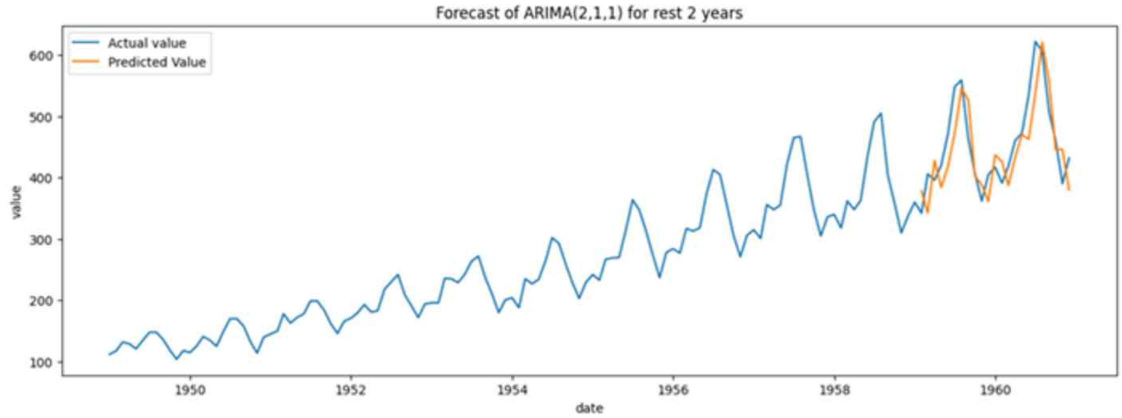


When examining the data, we can observe some slight gaps. Assuming a seasonality of 12 and plot the SARIMA model, resulting in the following.

2) SARIMA(1,1,2)(1,1,1,12)

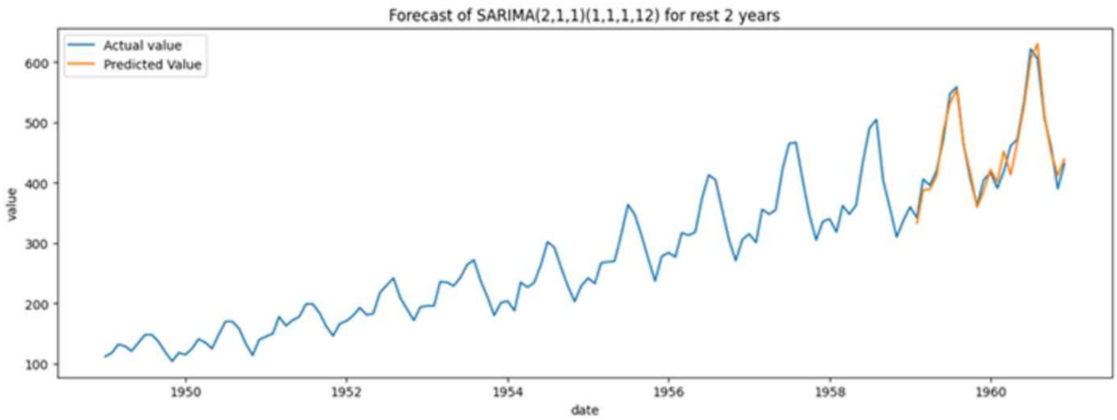


1) ARIMA(2,1,1)



When examining the data, we can observe some slight gaps. Assuming a seasonality of 12 and plot the SARIMA model, resulting in the following.

2) SARIMA(2,1,1)(1,1,1,12)

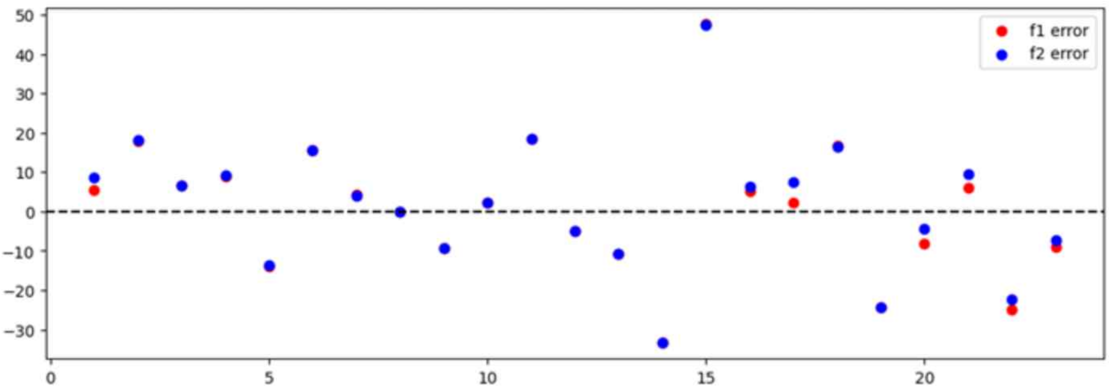


●Evaluate the best model in terms of estimation (best fit) and forecasting performance.

Actually I assume that ARIMA(2,1,1) will be the best fit model because it has statistically significant coefficient and small AIC, SBC.

Let model_1=ARIMA(1,1,2) and forecast of model_1=f1.

Let model_2=ARIMA(2,1,1) and forecast of model_2=f2.




```
Actual value:342, f1 forecast:336.49387677946066, f2 forecast:333.41680980774277
avg f1:455.4501, avg f2:454.5131
var of f1:5551.7343, var of f2:5548.3724
mean squared prediction error of f1: 285.8315, mean squared prediction error of
f2: 283.1435
```

1) Forecast the value of 1959_02

- f2 is close to the actual value and has less varaince and SSE than f1.

2) Test F test

- ARIMA(1,1,2)

```
<F test: F=0.9958498330597886, p=0.3861946722519276, df_denom=21, df_num=2>
Intercept    83.807224
f1            0.810897
dtype: float64
```

- ARIMA(2,1,1)

```
<F test: F=0.9941994049582115, p=0.3867773248490442, df_denom=21, df_num=2>
Intercept    85.812398
f2            0.808157
dtype: float64
```

- p-value of both model are less than 0.5, that mean we can reject the null hypothests and conclude both models are statistically significant.

3) Test GN-test

-> 0.46778828659720456

-Relationship between the two models is not statistically significant.

4) Test DM-test

-> 1.369073225479072

- Infer that either the first model (f1) or the second model (f2) outperforms the other in terms of forecasting performance."

Finally, I conclude that ARIMA(2,1,1) is the best fit moel