Forecasting American Airlines Stock Prices: A Time Series Approach

Rajrana Barman

May 2024

Summary

1 Problem Statement and Motivation

In this project our primary motive is to develop Time Series Model to forecast closing price of **American Airlines** Stock Price Dataset from the year 8th February 2013 to 7th February 2018. Here predicting the stock price can provide insights into market dynamics.

2 Understanding the data

	close
date	
2013-02-08	14.75
2013-02-11	14.46
2013-02-12	14.27
2013-02-13	14.66
2013-02-14	13.99

Figure 1: Snippet 1

We shall be using the following data for our analysis:

Our dataset consists of 1259 data points from the date 8th February 2013 to 7th February 2018.So in this dataset we use the date as index and we have closing price as feature. This is how the summary of the data looks like :

	close
count	1259.000000
mean	38.393252
std	10.957636
min	13.020000
25%	34.330000
50%	40.870000
75%	46.535000
max	58.470000

Figure 2: Snippet 2

- From the summary we can see that Mean < Median < Mode of the data. Hence the data is negatively skewed.
- About 50% of the data lies equal to or below 40 USD whereas the Maximum price is 58.47 USD

3 Visualizing the data

Figure 3: Snippet 3

- We have plot the data over time and visually inspected that there are patterns that change over time also the statistical properties of a dataset may not be constant over time.
- Nextly, we performed Train - Test split of the dataset. The training data consists from the date '2013-02-08' to '2017-02-07' and the test data consists from the date '2017-02-08' to '2018-02-07'.

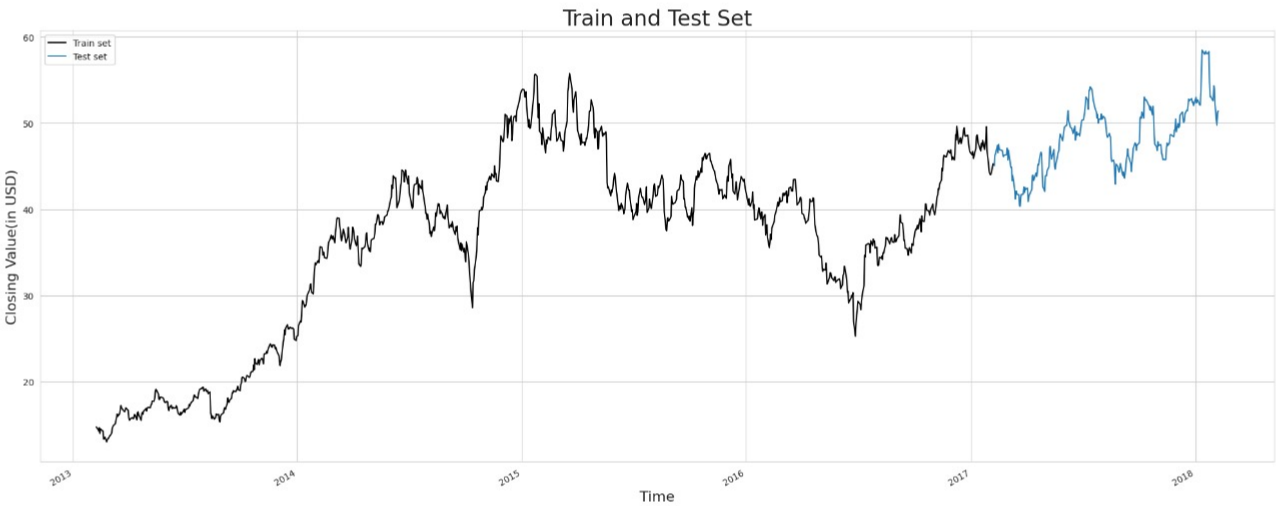

Figure 4: Snippet 4

- To determine correlation between each observation in the time series data with observations at previous time steps and also to detect model parameters we are plotting ACF and PACF.

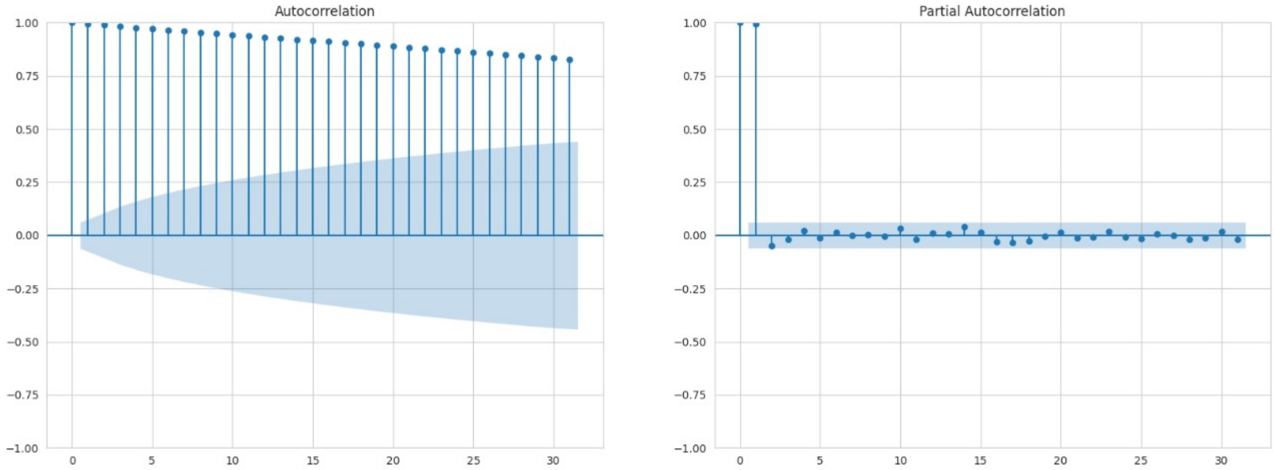

Figure 5: Snippet 5

- The ACF reveals that the autocorrelation doesn't tails off easily once again it indicates Non - Stationarity.
- The PACF reveals that the partial autocorrelation tails off after lag 1.
- So we perform 1st order differencing. This is how it looks like.

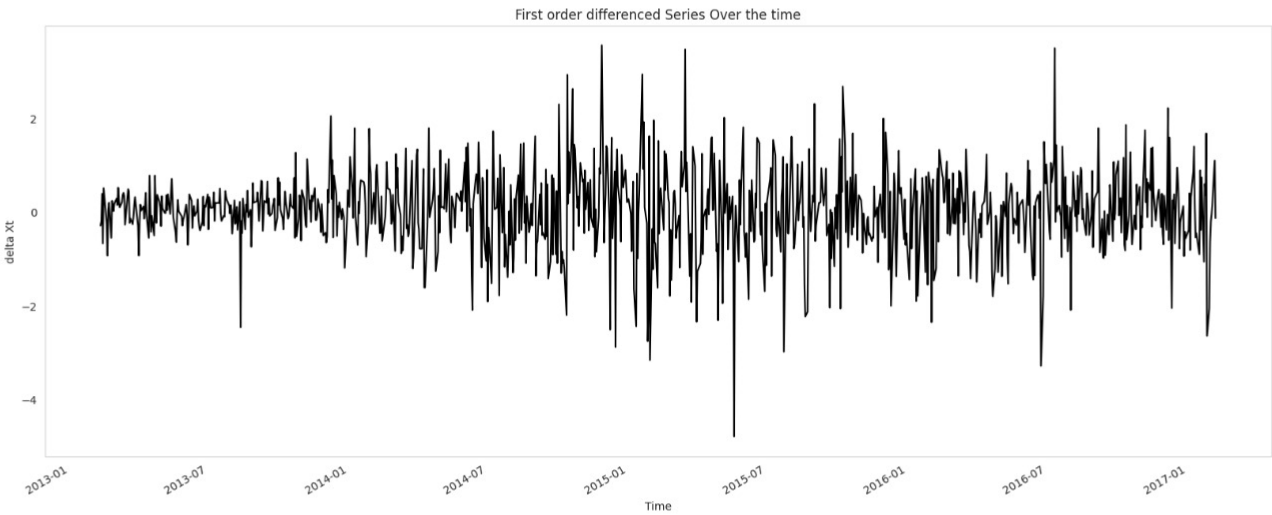

Figure 6: Snippet 6

- This is essentially calculating the difference between consecutive observations.
- The differenced series represent the change in closing price from one period to the next.
- We are effectively trying to remove trend component here. The delta Xt values are fluctuating around 0 hence the mean is 0 here and the variance remains constant overtime indicating stationarity.
- We then plot the ACF and PACF again:

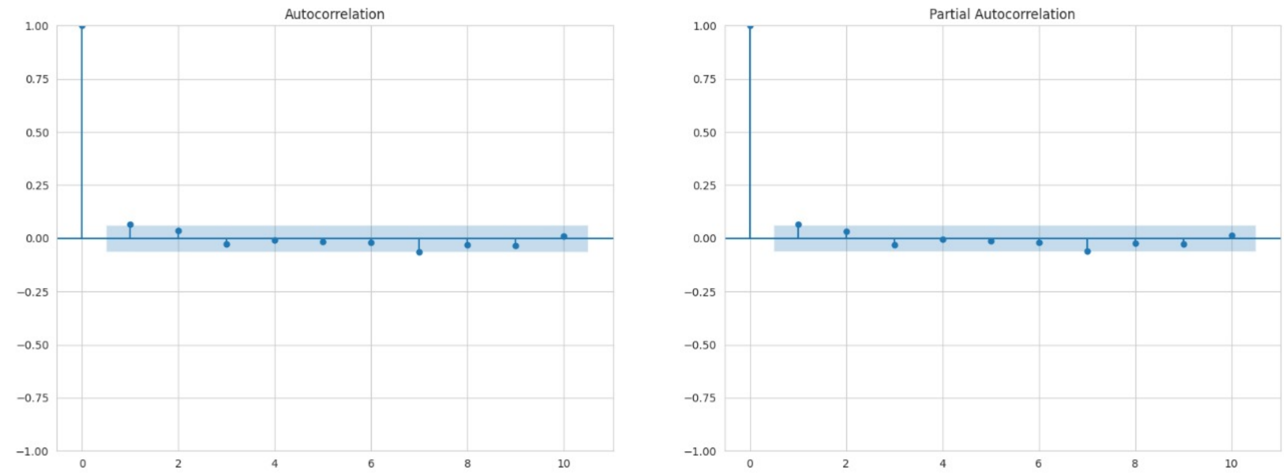

Figure 7: Snippet 7

4 Model 1

- The ACF tails off after lag 1. Hence we choose the parameter p as 1.
- The PACF tails off after lag 1. Hence we choose the parameter q as 1.
- Hence we fit ARIMA(1,1,1) model.The results are given below.

Dep. Variable:	close	**No. Observations:**	1006
Model:	SARIMAX(1, 1, 1)	**Log Likelihood**	-171.086
Date:	Fri, 03 May 2024	**AIC**	255.172
Time:	20:21:08	**BIC**	256.818
Sample:	0	**HQIC**	255.637
	- 1006		
Covariance Type:	opg		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	0.0224	0.024	0.934	0.035	-0.025	0.069
ar.L1	0.2738	0.356	0.769	0.044	-0.424	0.972
ma.L1	-0.2057	0.362	-0.568	0.057	-0.916	0.504
sigma2	0.7365	0.022	33.042	0.000	0.693	0.780

Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	0.665
Prob(Q):	0.92	**Prob(JB):**	0.76
Heteroskedasticity (H):	2.31	**Skew:**	-0.26
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.46

- The AIC and BIC of the model is 255.172 and 256.818 is moderate which indicates there may not be outliers or unusual patterns in the data that are not accounted for by the model.
- Now in the summary of the model we can see that Ljung Box test has p-value $= 0.92 > 0.05$ which indicates there is no significant autocorrelation in the residuals.
- The p-value of Jarque - Bera Test is $0.76 > 0.05$ hence we accept Null Hypothesis which tells that errors are normally distributed.

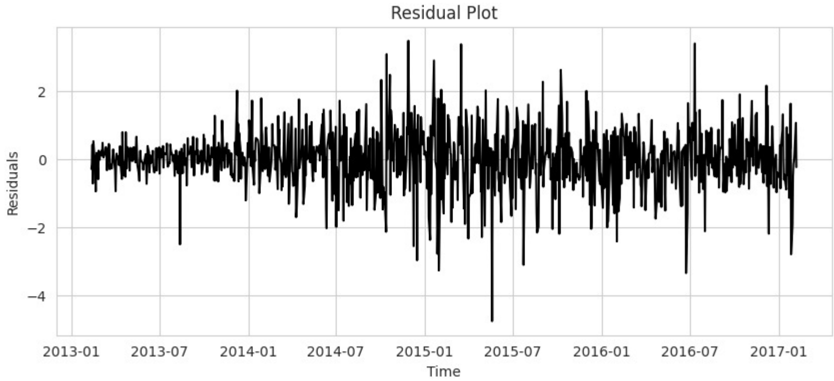

Figure 8: Snippet 8

- Again checking the residual plot we get that residuals have mean 0 and constant variance i.e., Homoscedasticity.
- We plot ACF and PACF to see whether the autocorrelation and partial autocorrelation of residual values are within the range of $(-2/\sqrt{n}, 2/\sqrt{n})$.

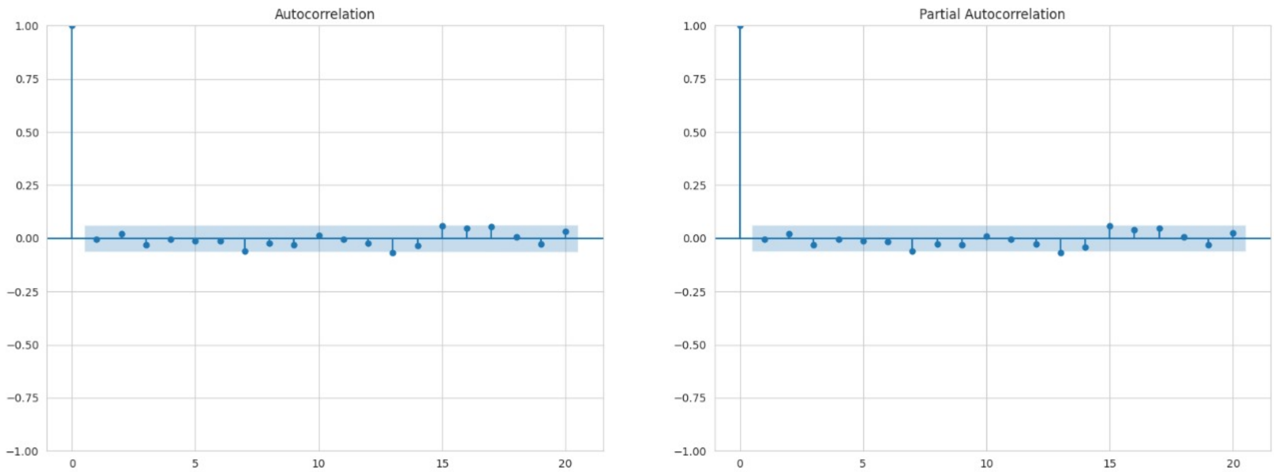

Figure 9: Snippet 9

- Checking the ACF and PACF plot we get that the autocorrelation and partial autocorrelation values are within the range of $(-2/\sqrt{n}, 2/\sqrt{n})$.
- After fitting our model on the test data we then move on to forecast closing prices using the test data. This is done to check the performance of our model on test set.

date	close	Predicted$_{v}alue$
2017-02-08	45.060000	45.205172
2017-02-09	46.300000	46.237158
2017-02-10	46.450000	46.268271
2017-02-13	47.410000	47.299145
2017-02-14	46.570000	46.329954

- The table shows the original and the forecasted closing prices. Using this we calculate the MAE and MAPE as a measure to check accuracy.
- MAPE = 1.01 and MAE = 1.05 which is quite low indicating a good fit.

5 Model 2

- Since ACF and PACF both tails off after lag 1 so we will see whether it tails off or not. So in this regard we fit another model by taking the parameters p = 2, q = 2 and d remaining the same.

- Hence we fit ARIMA(2,1,2) model.The results are given below.

Dep. Variable:	close	**No. Observations:**	1006
Model:	SARIMAX(2, 1, 2)	**Log Likelihood**	-269.382
Date:	Fri, 03 May 2024	**AIC**	289.764
Time:	21:28:58	**BIC**	280.229
Sample:	0	**HQIC**	261.961
	- 1006		
Covariance Type:	opg		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	0.0443	0.047	0.948	0.0343	-0.047	0.136
ar.L1	-0.4640	0.453	-1.024	0.0306	-0.0352	0.424
ar.L2	0.0031	0.364	0.009	0.009	-0.0710	0.717
ma.L1	0.5302	0.453	1.170	0.0242	-0.0358	1.418
ma.L2	0.0677	0.356	0.190	0.0045	-0.0629	0.765
sigma2	0.7356	0.022	33.033	0.000	0.0692	0.779

Ljung-Box (L1) (Q):	0.006	**Jarque-Bera (JB):**	4.13
Prob(Q):	0.99	**Prob(JB):**	0.68
Heteroskedasticity (H):	2.28	**Skew:**	-0.25
Prob(H) (two-sided):	0.00	**Kurtosis:**	5.46

- The AIC and BIC of the model is 289.764 and 280.229 is moderate which indicates there may not be outliers or unusual patterns in the data that are not accounted for by the model but these values are greater than Model 1.
- Now in the summary of the model we can see that Ljung Box test has p-value $= 0.99 > 0.05$ which indicates there is no significant autocorrelation in the residuals.
- The p-value of Jarque - Bera Test is $0.68 > 0.05$ hence we accept Null Hypothesis which tells that errors are normally distributed.

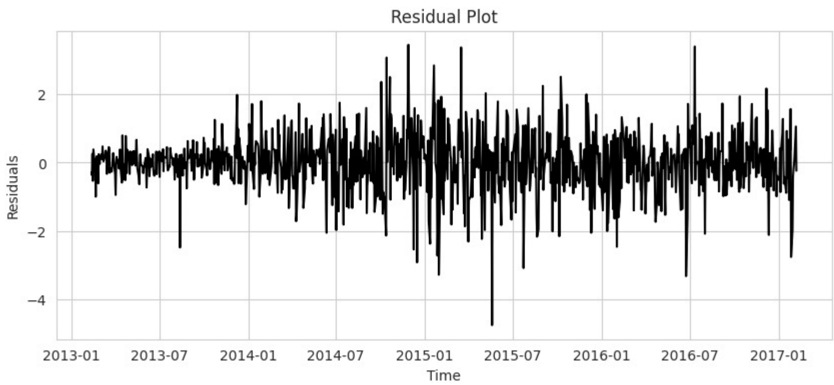

Figure 10: Snippet 10

- Again checking the residual plot we get that residuals have more upper and lower fluctuations around the value 0 as compared to model 1.
- We plot ACF and PACF to see whether the autocorrelation and partial autocorrelation of residual values are within the range of $(-2/\sqrt{n}, 2/\sqrt{n})$.

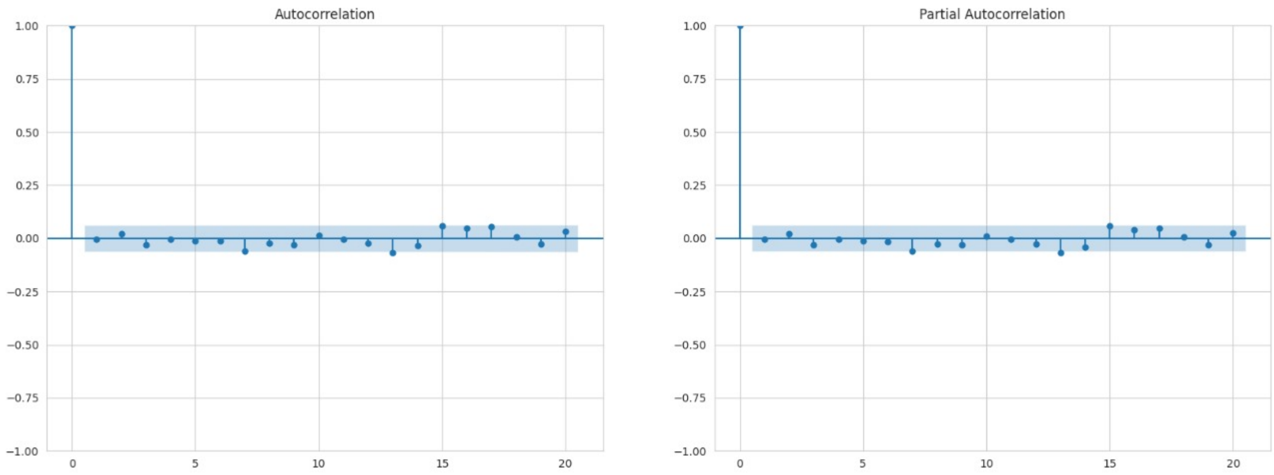

Figure 11: Snippet 11

- Checking the ACF and PACF plot we get that the autocorrelation and partial autocorrelation values are within the range of $(-2/\sqrt{n}, 2/\sqrt{n})$.
- After fitting our model on the test data we then move on to forecast closing prices using the test data. This is done to check the performance of our model on test set.

date	close	Predicted$_{v}alue2$
2017-02-08	45.060000	45.405172
2017-02-09	46.300000	46.137158
2017-02-10	46.450000	46.218271
2017-02-13	47.410000	47.199145
2017-02-14	46.570000	46.229954

- The table shows the original and the forecasted closing prices. Using this we calculate the MAE and MAPE as a measure to check accuracy.
- MAPE = 2.11 and MAE = 1.91 which is quite low indicating a good fit.

6 Conclusion

- Comparing the Model 1 and Model 2 we get the MAP and MAPE values for Model 1 is lesser indicating that Model 1 is a better fit.
- Also the AIC and BIC values for Model 1 is less this is one more reason to conclude that Model 1 is a better fit in this case.

7 Decomposition

7.1 Multiplicative

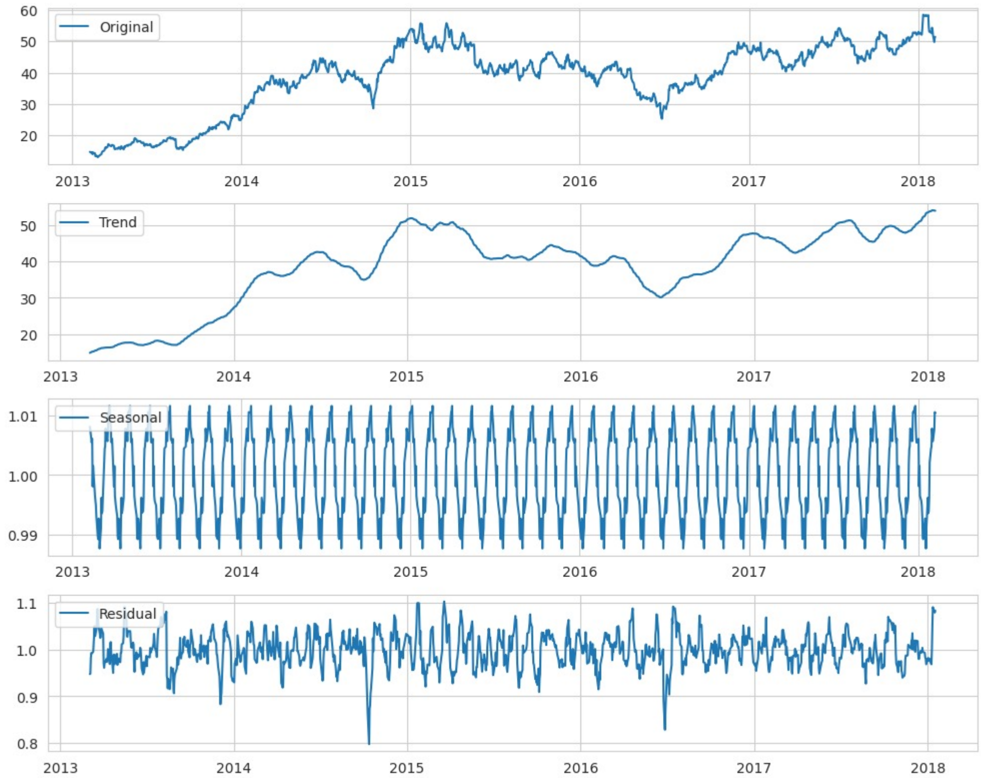

Figure 12: Snippet 12

- We have decomposed the into 3 components i.e.,Trend component, Seasonal and error component.
- The trend is increasing in nature identifying overall growth of closing prices.
- The seaonal component has a period of 7 assessing the seasonality is stable.
- The residual component represents the variability in the data that is not explained by the trend or seasonality.