What we can learn from tourists visiting Taiwan before and after pandemic?

Time series forecasting using ARIMA and PyCaret

Link Yang
3/12/2024

Outline

- Introduction
- Executive Summary
- Exploratory Data Analysis
- Modeling Preparation and Results
- Conclusion and Future Direction

Introduction

Background

- Taiwan's travel and tourism industry plays a significant role in its economy,
- The foreign exchange earnings from tourism for the year 2019 accounted for **4.43% of Taiwan's GDP.**
- The total number of visitors to Taiwan in 2019 reached an **all-time-high record, 11.85 million.**

After the outbreak of COVID-19, the total number of tourists was less than 0.14 million in 2021.

Determine the trend in the influx of visitors before pandemic

&

Build a time series forecasting model

Executive Summary

- We validated the continuous growth in visitors to Taiwan since 2011.
- We successfully construct an **ARIMA (0, 1, 1)** model with satisfactory fitness.
- We leveraged PyCaret to further develop two predictive models that effectively leverage past data to make realistic forecasting.
    - **Seasonal ARIMA (SARIMA) model**
    - **Exponential smoothing model**

Exploratory Data Analysis

Data Source

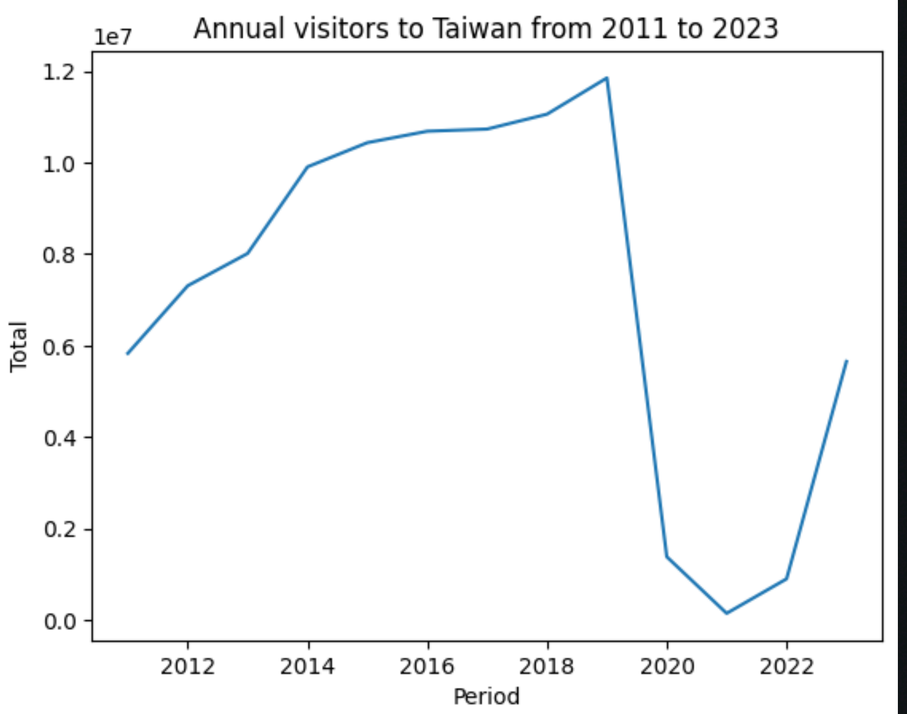
- A single CSV file downloaded from the official website of Taiwan Tourism Bureau.
- It contains the monthly number of visitors to Taiwan by their gender, residence and purpose from January 2011 to November 2023.
- All of the visitors were categorized into 8 groups based on their purpose of travel.

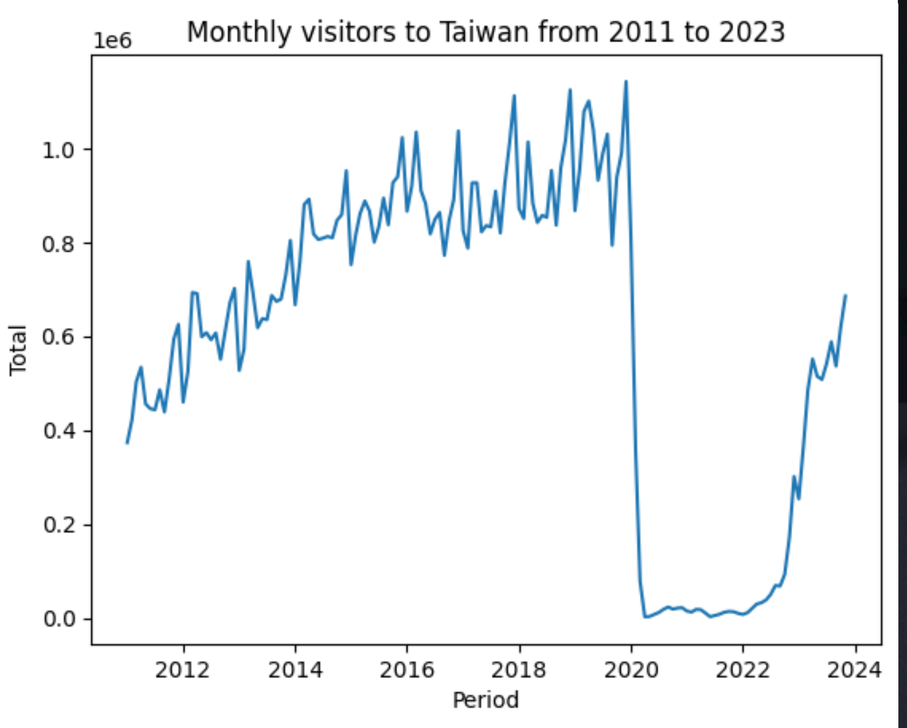
https://eng.taiwan.net.tw/

Exploratory Data Analysis

Is there any trend and seasonal fluctuation?

- The influx of visitors to Taiwan has exhibited an upward trajectory since 2011.
- It culminated in a record high of 11.85 million in 2019.
- There is a seasonal pattern in the total number of visitors.
- This pattern was only evident before the outburst of pandemic.

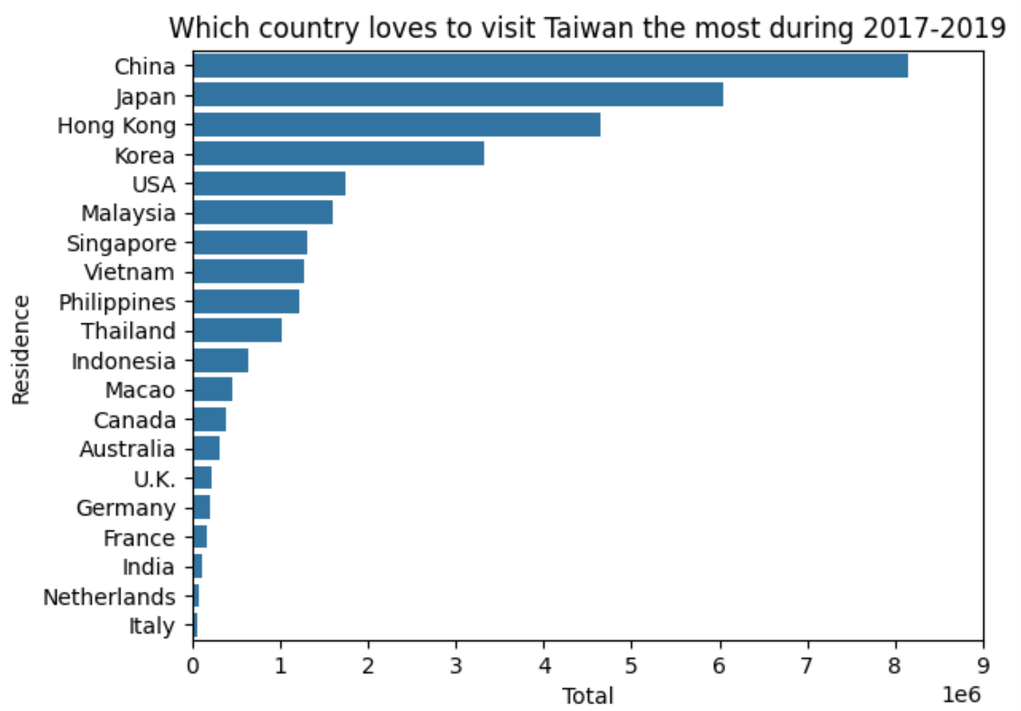


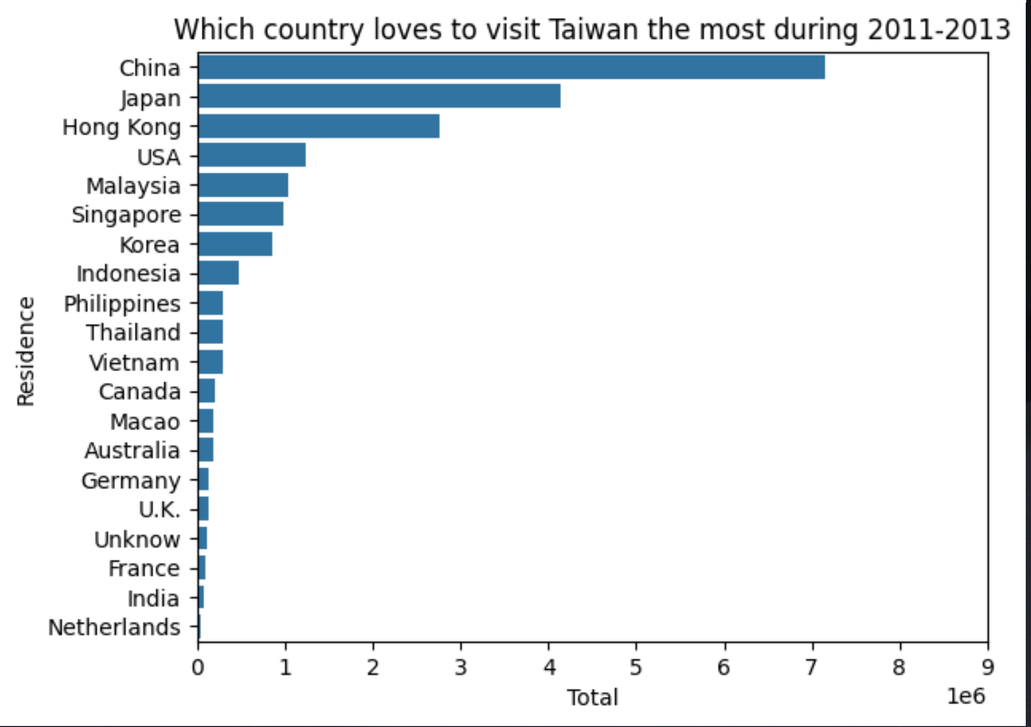


Exploratory Data Analysis

Which country loves to visit Taiwan the most?

- Mainland China loves to visit Taiwan the most during 2017-2019, followed by Japan, Hong Kong, Korea, and USA.
- The ranking of the number from earlier time points, 2011-2013, is mostly similar to 2017-2019.
- Some countries have more drastic change in the total number now and then.

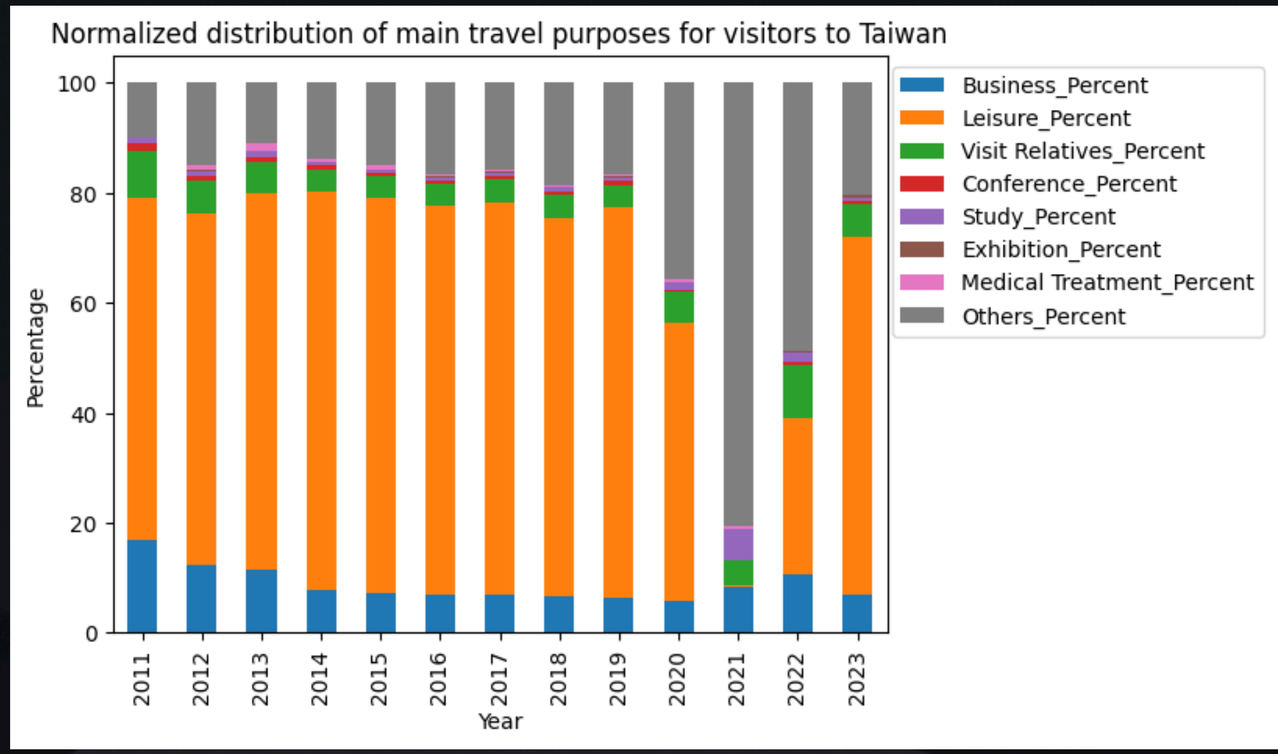




Exploratory Data Analysis

What was the main travel purpose ?

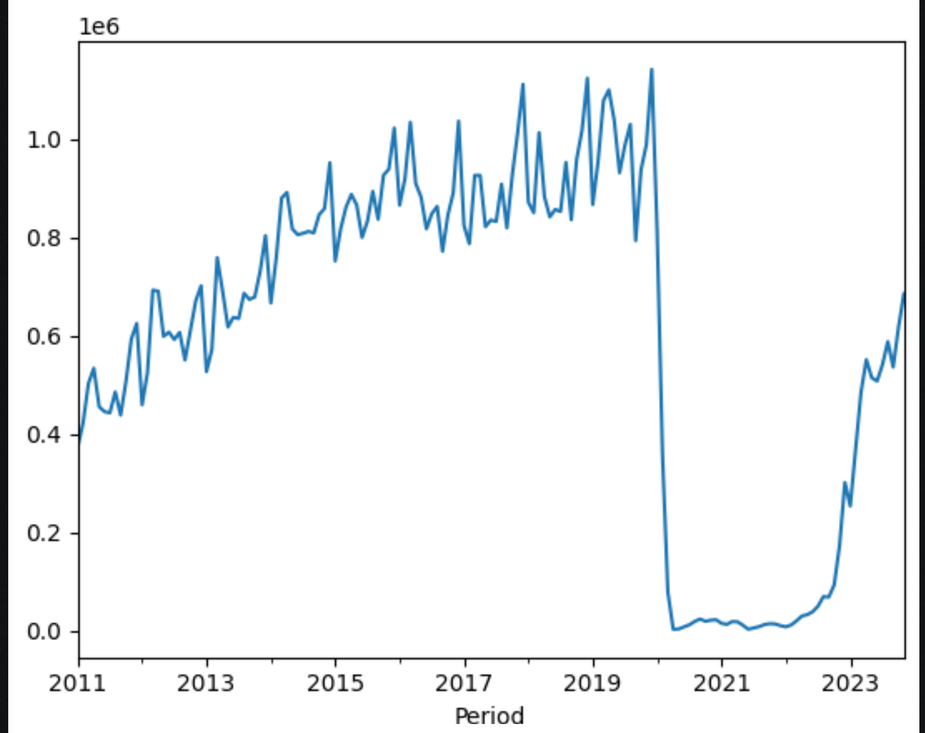
- Majority of visitors to Taiwan were for leisure or recreational purposes.
- Overall the distribution of travel purpose to Taiwan remained similar throughout the decade.

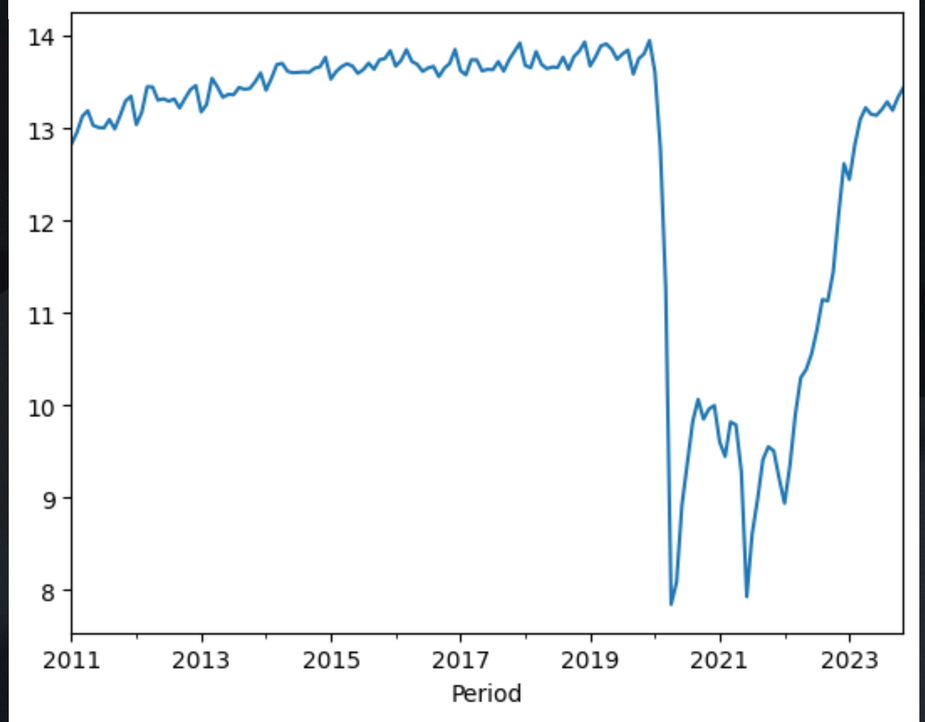


Modeling Preparation

Assess stationarity

- Original data
  - **The mean and variance increase over time.**
  - KPSS *p*-value < 0.05 ➜ non-stationary
  - Dicky-Fuller *p*-value > 0.05 ➜ non-stationary
- After natural log transformation
  - **Only the variance becomes more constant.**
  - KPSS *p*-value < 0.05 ➜ non-stationary
-

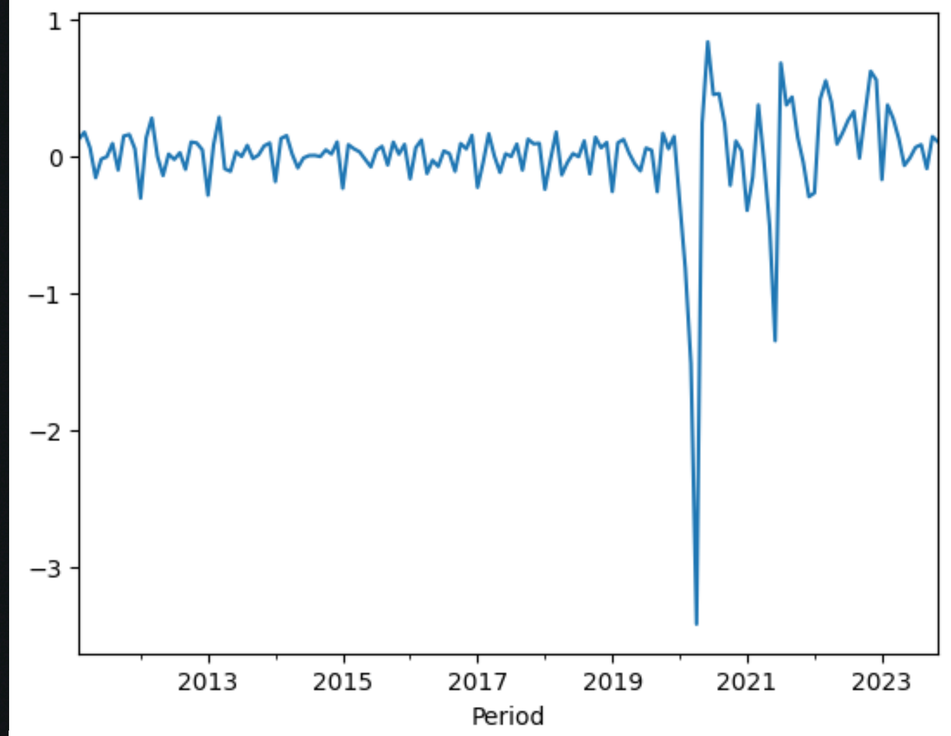


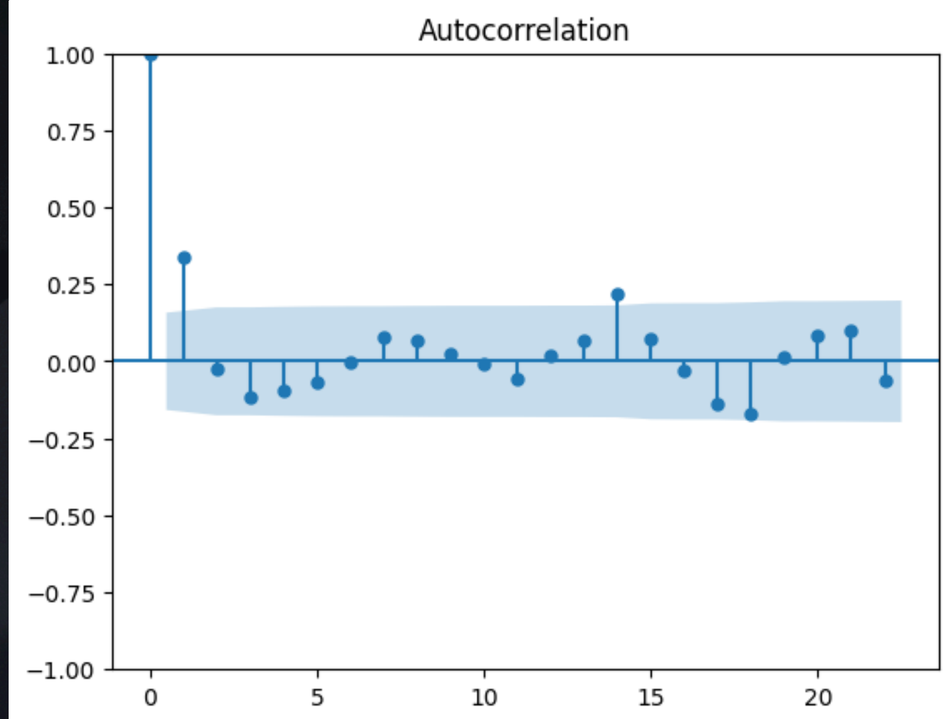


Modeling Preparation

Assess stationarity & autocorrelation

- After log transformation & differencing
  - KPSS *p*-value > 0.05 ➜ stationary
  - Dicky-Fuller *p*-value < 0.05 ➜ stationary
  - **Both tests indicate stationarity, the series can be deemed to be trend-stationary.**
- Plot Autocorrelation Function (ACF)
  - **Autocorrelation is detected!**





Modeling Preparation

Optimize ARIMA model parameters

- 1st Approach
  - **Root mean squared error (RMSE)**
  - **Mean absolute error (MAE)**
- 2nd Approach
  - **Akaike Information Criterion (AIC)**
  - **Bayesian Information Criterion (BIC)**
- MAE, AIC and BIC all favor the same model **ARIMA(0, 1, 1).**

	p	d	q	RMSE	MAE
0	1	0	1	0.391368	0.278261
1	2	1	0	0.391819	0.274175
2	1	0	2	0.392018	0.275629
3	1	1	2	0.392148	0.280461
4	0	1	1	0.392511	0.271685
5	2	0	1	0.393123	0.277563

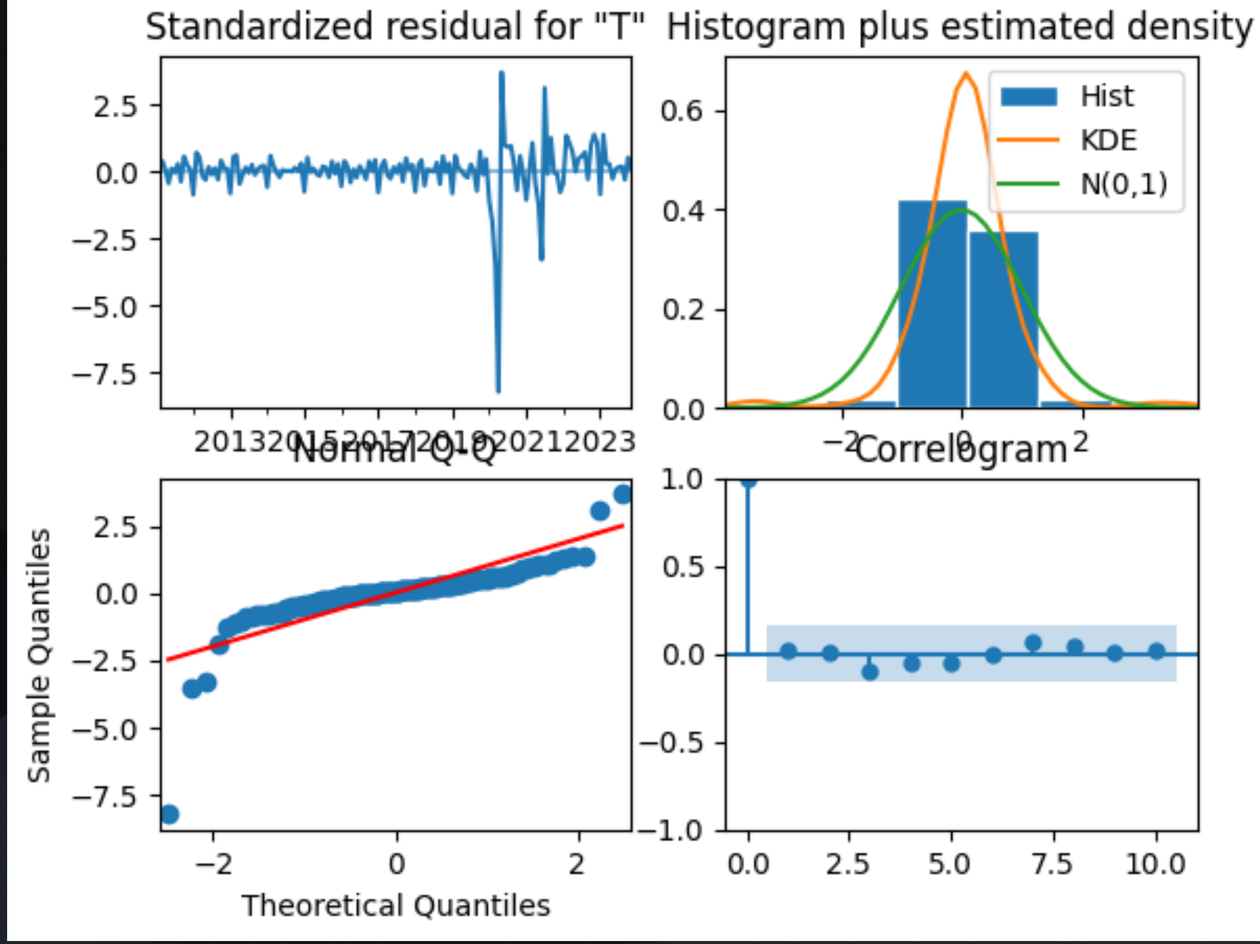
	p	d	q	AIC	BIC
0	0	1	1	126.352402	132.426307
1	1	1	0	128.740083	134.813988
2	2	1	0	127.049990	136.160847
3	0	1	2	128.064826	137.175684
4	1	1	1	128.162109	137.272966
5	2	1	1	128.188567	140.336377

Modeling Results

Residual evaluation

- L1 test *p*-value > 0.05 ➜ **no autocorrelation** in the residuals of the time series
- JB test *p*-value = 0 ➜ the residual distribution **significantly deviates from a normal distribution**
- The histogram and Q-Q plots both show the residuals are not normally distributed.
- The deviation was caused by the onset of global pandemic in early 2020.

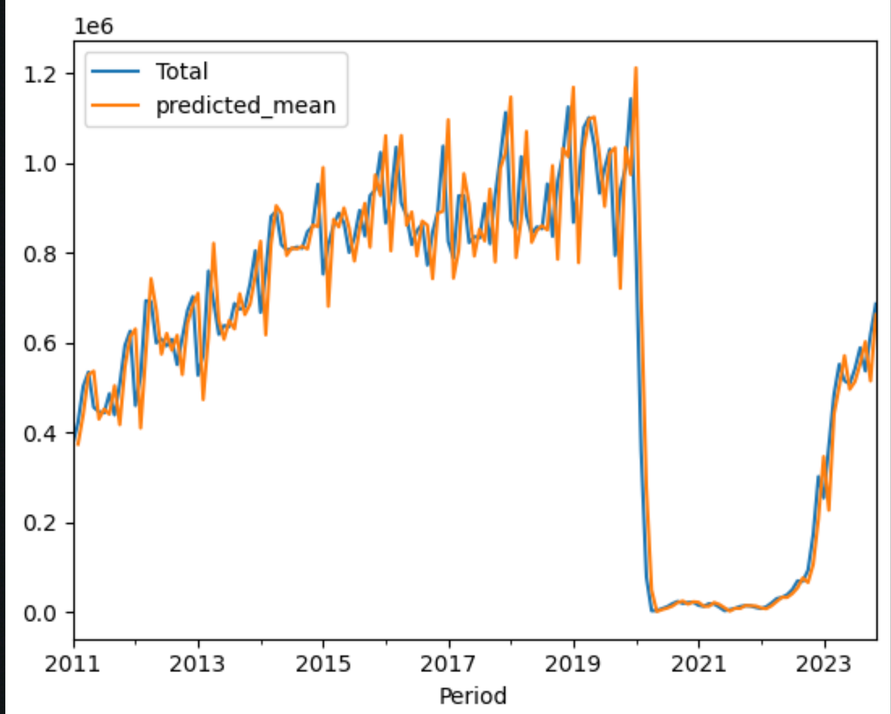
Ljung-Box (L1) (Q):	0.04	**Jarque-Bera (JB):**	6311.40
Prob(Q):	0.83	**Prob(JB):**	0.00
Heteroskedasticity (H):	21.79	**Skew:**	-3.61
Prob(H) (two-sided):	0.00	**Kurtosis:**	33.52

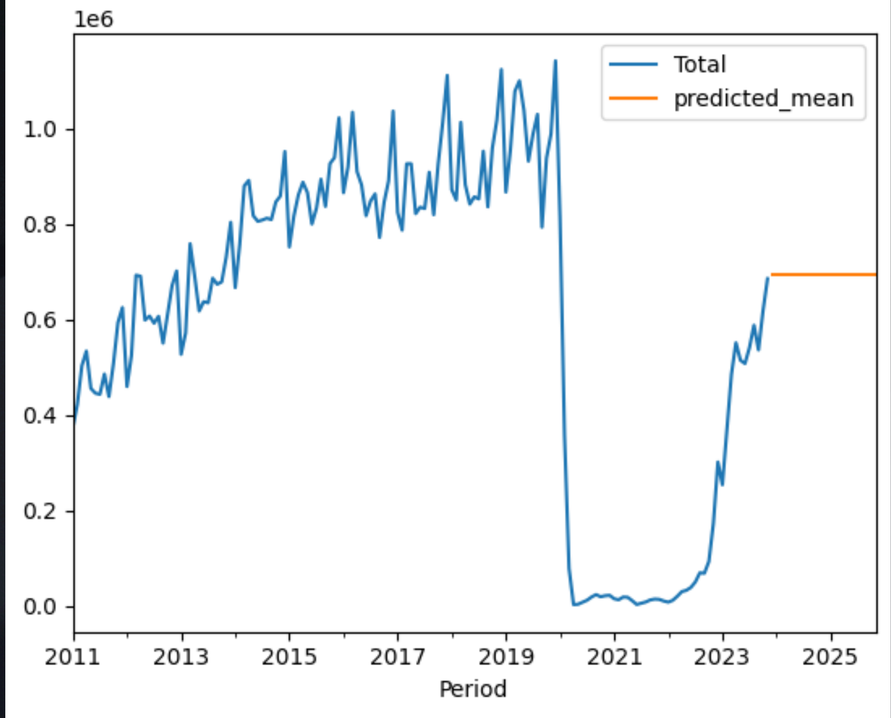


Modeling Results

Visual inspection & forecasting

- Our ARIMA(0, 1, 1) model fits tightly to our existing data.
- However, forecasts more than one period in the future are simply equal to the first forecast of the sample.
- **Overall fitting is good but forecasting is less informative** ➜ Any alternatives?



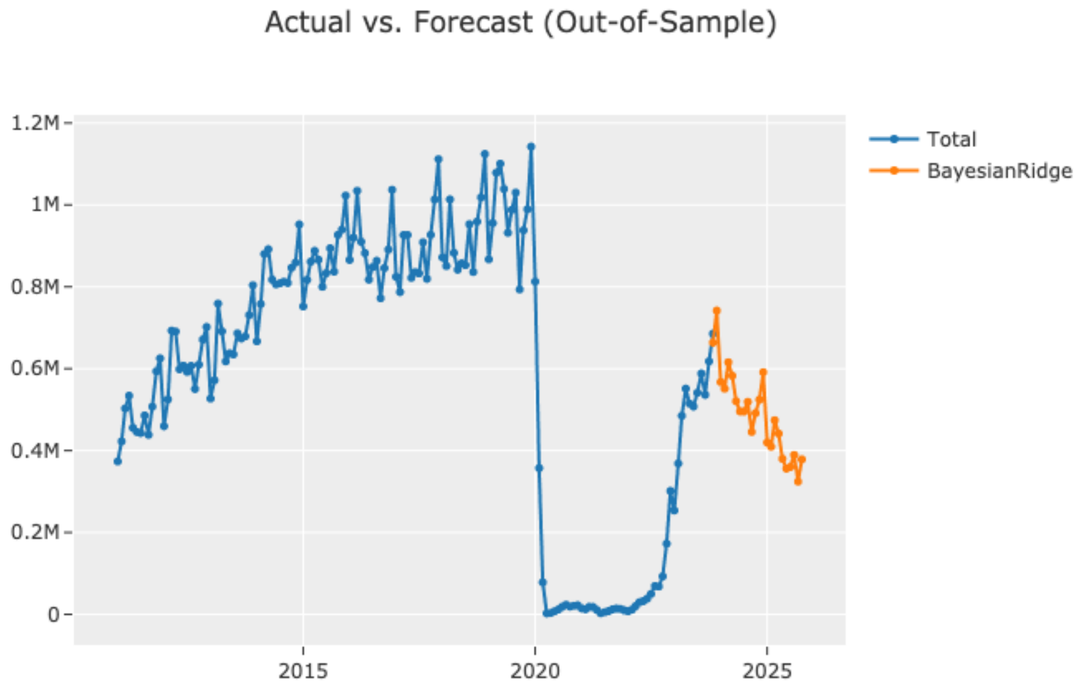


Modeling Results

Use PyCaret to explore other solutions

- **Bayesian Ridge model** highlighted at the top exhibits the lowest error scores and outperforms all other models.
- Surprisingly, the model **predicts another downturn in the next 24 months**, which contradicts our expectation and common scenario.
- Most top performers in this preliminary experiment yield a similar downtrend.

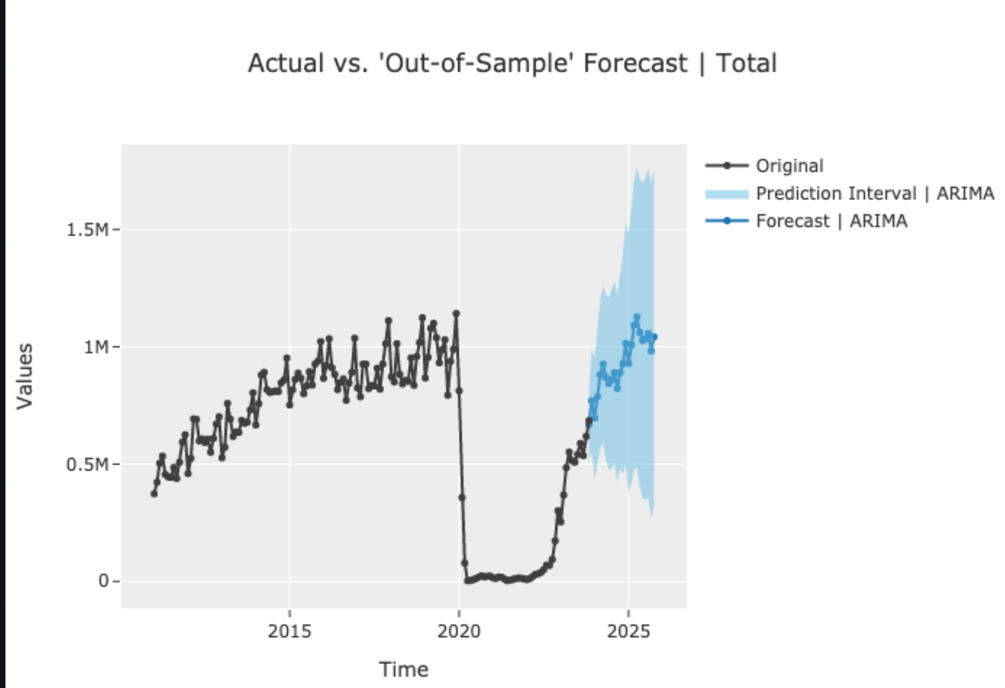
Model	MASE	RMSSE	MAE	RMSE
Bayesian Ridge w/ Cond. Deseasonalize & Detrending	0.0763	0.0412	12750.7357	12750.7357
Lasso Least Angular Regressor w/ Cond. Deseasonalize & Detrending	0.0819	0.0443	13706.0791	13706.0791
Lasso w/ Cond. Deseasonalize & Detrending	0.0819	0.0443	13706.0819	13706.0819
Linear w/ Cond. Deseasonalize & Detrending	0.0819	0.0443	13706.0792	13706.0792
Elastic Net w/ Cond. Deseasonalize & Detrending	0.0819	0.0443	13706.0819	13706.0819
Ridge w/ Cond. Deseasonalize & Detrending	0.0819	0.0443	13706.0792	13706.0792

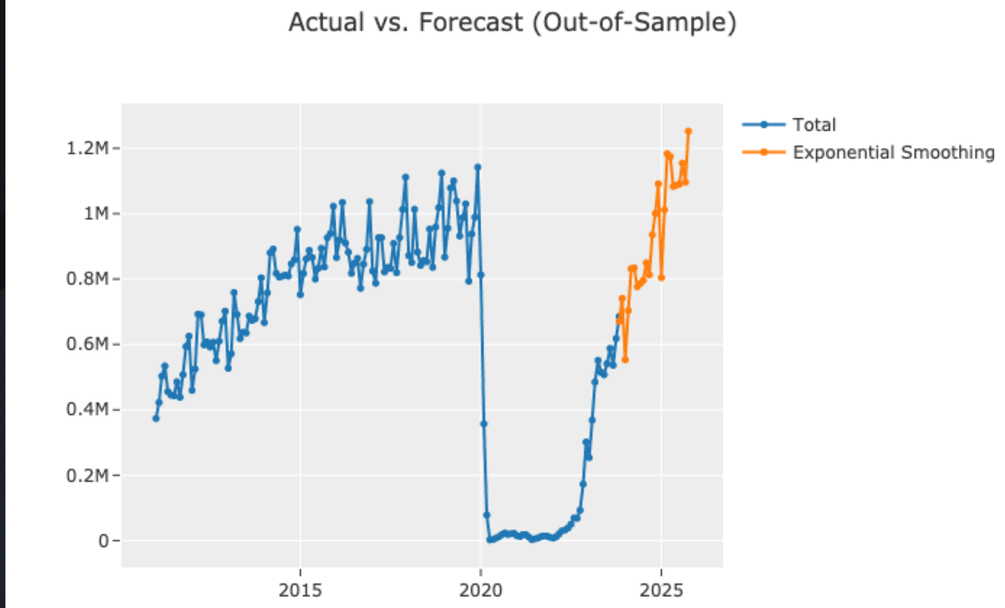


Modeling Results

Use PyCaret to explore other solutions

- Both the **SARIMA and Exponential Smoothing models** anticipate an upward trend over the next 24 months.
- This forecast appears to be more plausible than that generated by the Bayesian Ridge model, despite a slightly higher error score.





Conclusion

- We confirmed the **continuous growth in visitors to Taiwan since 2011**. Despite the unforeseen and unprecedented decline in the flux of tourists due to Covid-19, we were able to **successfully construct an ARIMA** model with decent fitness.

- Furthermore, we utilized PyCaret to develop two predictive models that effectively leverage past data to make realistic forecasting.

- The key takeaway is that selecting the best time series model cannot be solely determined by performance metrics. In addition to **model fitness**, ensuring the **resulting forecast is meaningful and sensible** is equally crucial.

Future Direction

- We could further explore recurrent neural network (RNN) and other deep learning models for time series forecasting.


Thank you!