1. **Understanding the Data**
   a) Identify the most and least trafficked routes
   Some of the most trafficked routes included the following (Jan 1985 - June 1989):
     - Sydney and Auckland (2.96 million total passengers, 126.71k tonnes of freight, 3,280.77 tonnes of mail)
     - Sydney and Singapore (1.44 million total passengers, 69.31k tonnes of freight, 1,147.66 tonnes of mail)
     - Sydney and Tokyo (1.29 million total passengers, 79.16k tonnes of freight, 2,423.64 tonnes of mail)
     - Sydney and Hong Kong (1.15 million total passengers, 51.61k tonnes of freight, 494.55 tonnes of mail)
     - Perth and Singapore (952.93k total passengers, 53.57k tonnes of freight, 383.26 tonnes of mail)

   Some of the least trafficked routes included:
     - Brisbane and Chicago (0 total passengers, 0 tonnes of freight, 0.009 tonnes of mail)
     - Adelaide and Harare (0 total passengers, 0 tonnes of freight, 0 tonnes of mail)
     - Brisbane and Colombo (0 total passengers, 0.2 tonnes of freight, 0 tonnes of mail)
     - Perth and Bandar Seri Begawan (0 total passengers, 149.46 tonnes of freight, 0 tonnes of mail)
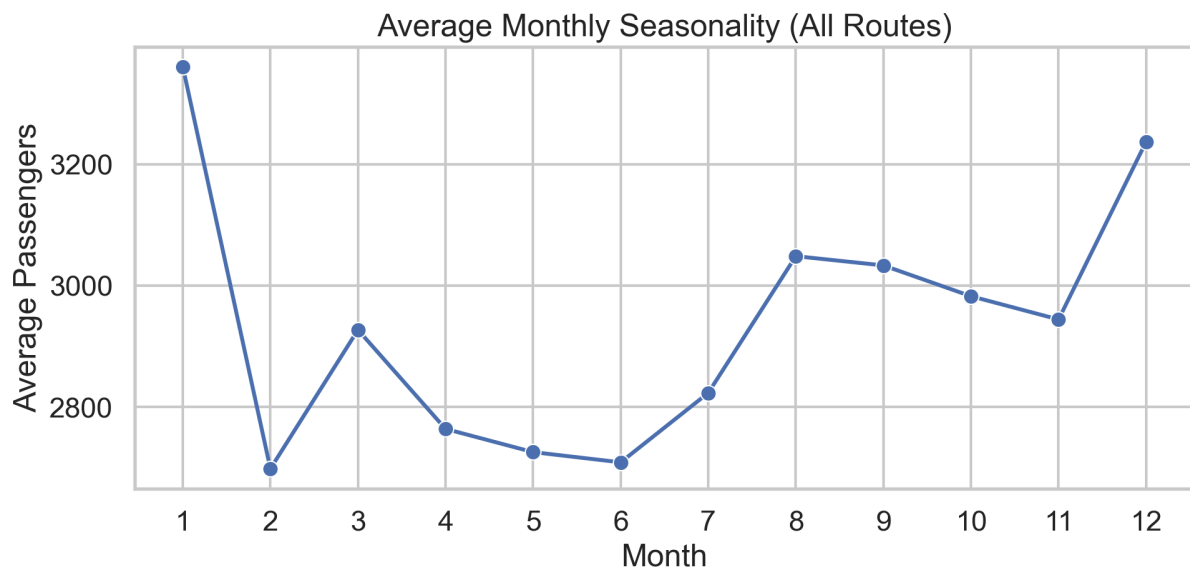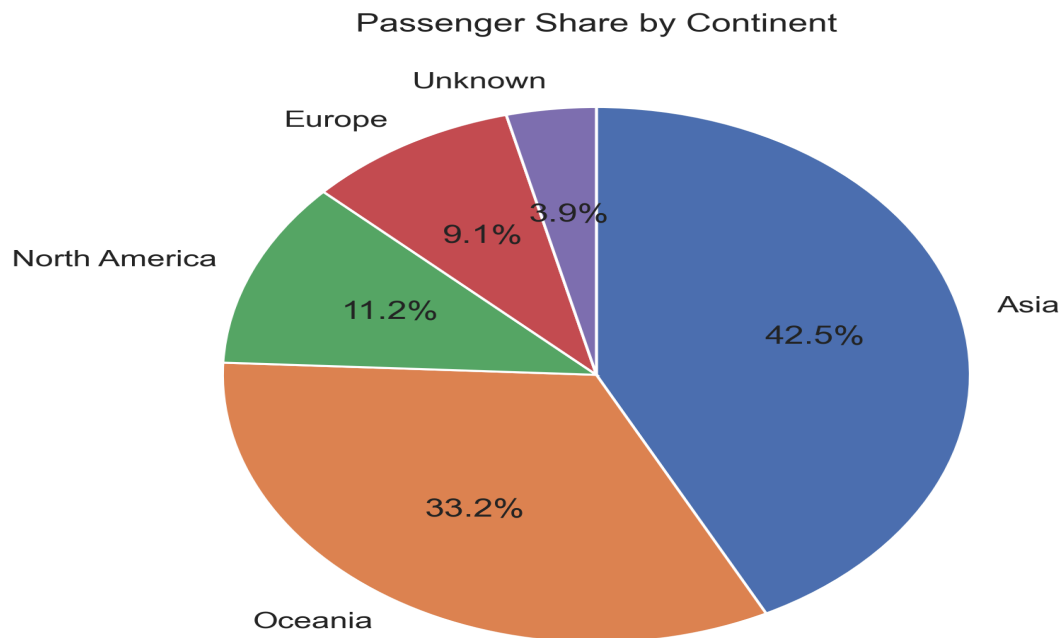   - Melbourne and Denver (0 total passengers, 0.054 tonnes of freight, 0 tonnes of mail)

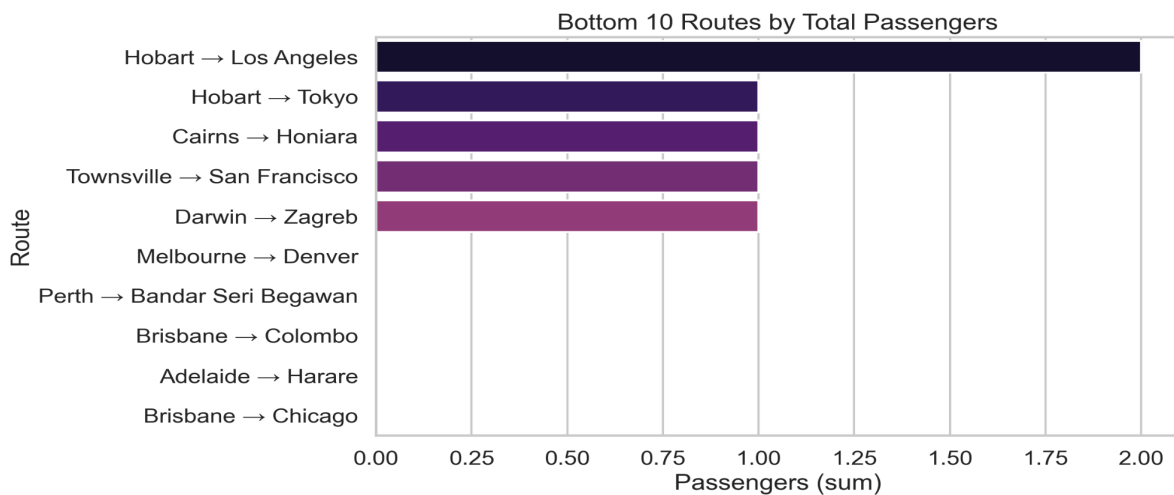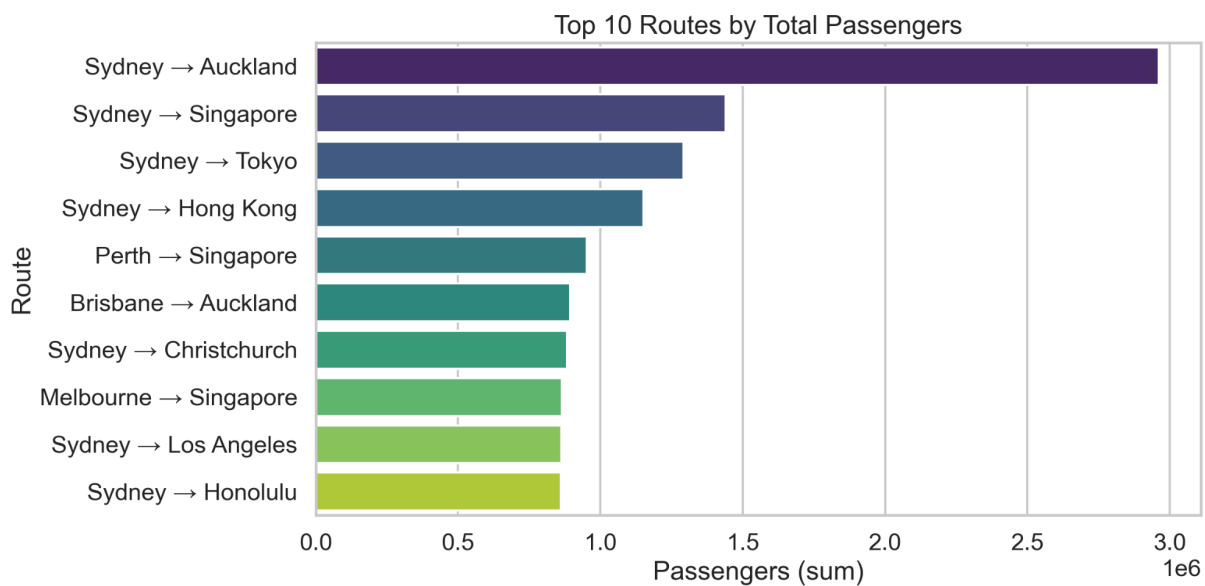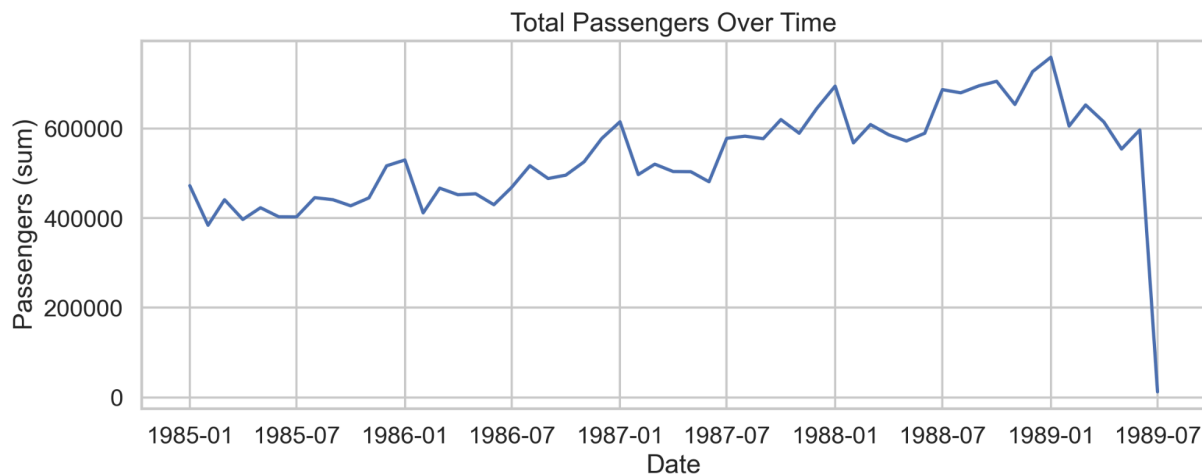   **b) Analyze trends and/or geographical patterns**
   I noticed that most of the passenger traffic came from Asia and Oceania, continents that are close in proximity to Australia. Due to this, some of the most popular foreign ports included Singapore, Auckland, Tokyo, Hong Kong, and Christchurch. Additionally, across all the Australian port passenger totals, I found that months December and January were the most busiest due to the holidays and weather in Australia, making it an extremely great tourist location then. Additionally, lots of boxing day (day after Christmas) sports games happen in this area, so many tourists love to travel for those occasions. Most of the time, by February, we usually observe a local minimum (hype dips), then in March it spikes again slightly for it to stagnate usually until around July. The pattern here stays somewhat consistent across the 4.5 years of data given, until we reach 1989, where the local minimums and maximums aren't reflective of the pattern observed in the previous 4 years.

   Looking at some of the city pairs, I realized that even though Sydney and Auckland had high passenger traffic coming in and out of it, the peaks and valleys were much more inconsistent compared to other Australian ports and Singapore. Routes connected to Singapore, for example, displayed steadier growth patterns with smaller fluctuations, while Sydney and Auckland tended to swing up and down with sharper highs and lows. This suggests that while
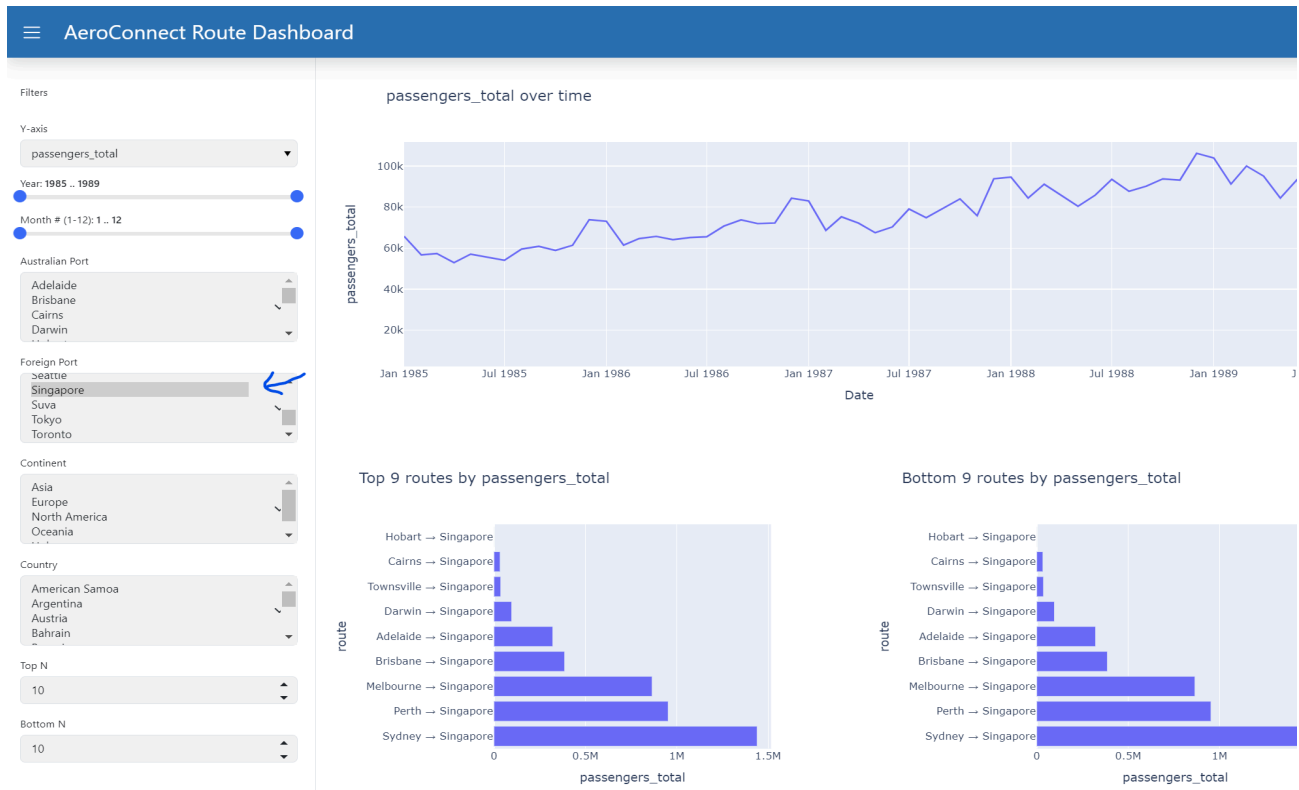
these two hubs were major traffic drivers, their demand was more volatile, whereas other ports maintained a more stable passenger flow.

c) Create visualizations to demonstrate trends & patterns determined in part b

Passenger Share by Continent

Average Monthly Seasonality (All Routes)

Total Passengers Over Time

Top 10 Routes by Total Passengers

Sydney → Auckland
Sydney → Singapore
Sydney → Tokyo
Sydney → Hong Kong
Perth → Singapore
Brisbane → Auckland
Sydney → Christchurch
Melbourne → Singapore
Sydney → Los Angeles
Sydney → Honolulu

Bottom 10 Routes by Total Passengers

Hobart → Los Angeles
Hobart → Tokyo
Cairns → Honiara
Townsville → San Francisco
Darwin → Zagreb
Melbourne → Denver
Perth → Bandar Seri Begawan
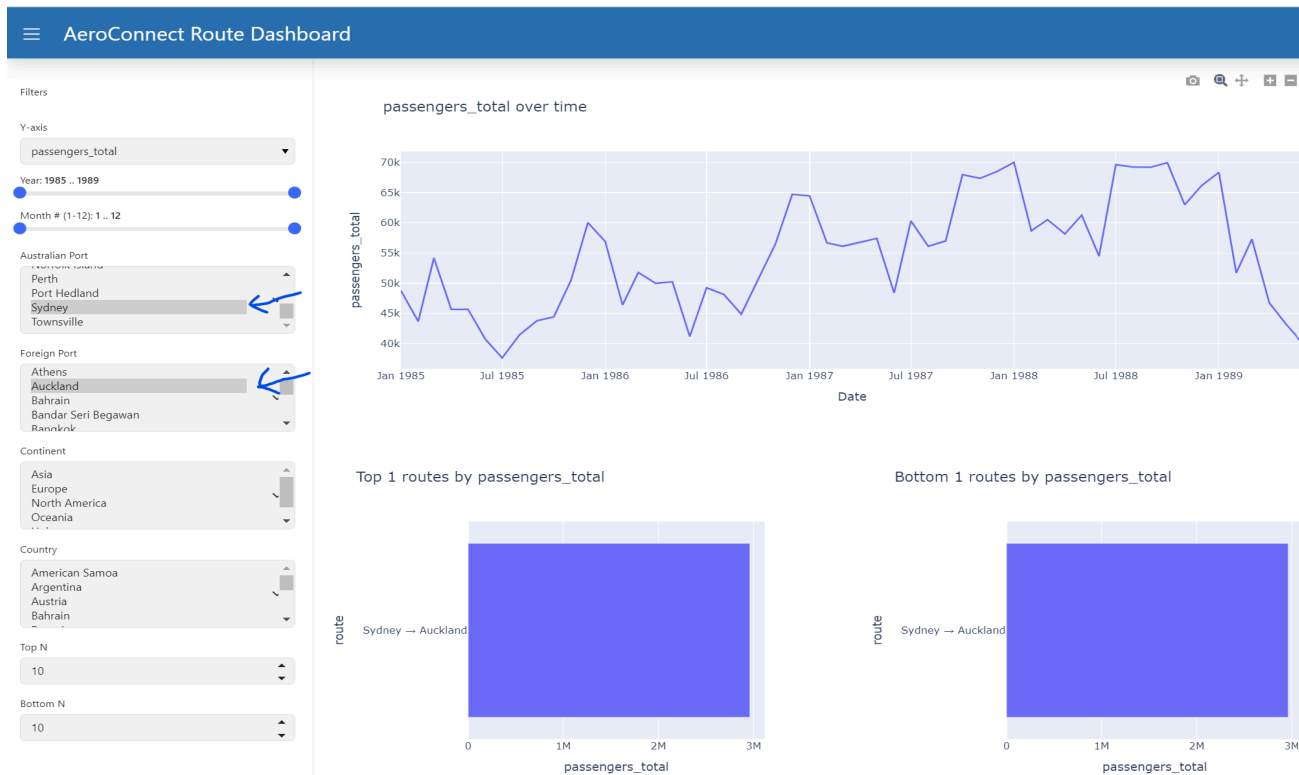Brisbane → Colombo
Adelaide → Harare
Brisbane → Chicago

From my Dashboard:

(Shows passenger totals overtime when foreign port is just Singapore)

(Shows passenger totals overtime when australian port is Sydney and foreign port is Auckland)

2. Build a Model

a) Your model should predict passenger traffic for the next 6–12 months on at least 1 city pair

Route	Predicted Values

Sydney → Singapore	```date,predicted_passengers```
	```1989-07-01,32214.94978365379```
	```1989-08-01,29946.52182835597```
	```1989-09-01,30914.875699718177```
	```1989-10-01,32144.011213272002```
	```1989-11-01,32119.40948811096```
	```1989-12-01,33271.43043930177```
	```1990-01-01,35203.398531082596```
	```1990-02-01,31125.65066205877```
	```1990-03-01,34938.40685201779```
	```1990-04-01,31928.49870923276```
	```1990-05-01,28255.009916558243```
	```1990-06-01,30576.39882423147```

Perth → Singapore	```date,predicted_passengers```
	```1989-07-01,23635.24694511035```
	```1989-08-01,22119.1472940862```
	```1989-09-01,22973.9846010786```
	```1989-10-01,23884.19797783349```
	```1989-11-01,22894.209824491398```
	```1989-12-01,30932.616970102787```
	```1990-01-01,26957.72063424316```
	```1990-02-01,24524.268311174546```
	```1990-03-01,25098.250808383484```
	```1990-04-01,24136.848722573646```
	```1990-05-01,21573.40742096973```
	```1990-06-01,24461.113379907718```

Melbourne → Singapore

```
date,predicted_passengers
1989-07-01,19586.15555620497
1989-08-01,18957.834204548155
1989-09-01,19874.15192767809
1989-10-01,18931.6935209761
1989-11-01,19405.4885926266
1989-12-01,23314.531613760228
1990-01-01,23807.06062815905
1990-02-01,20679.79186770392
1990-03-01,21935.519381369024
1990-04-01,20524.813789733027
1990-05-01,18329.285088366396
1990-06-01,20482.91493940039
```


Sydney → Singapore: Actual vs Predicted Passengers

Perth → Singapore: Actual vs Predicted Passengers

Melbourne → Singapore: Actual vs Predicted Passengers
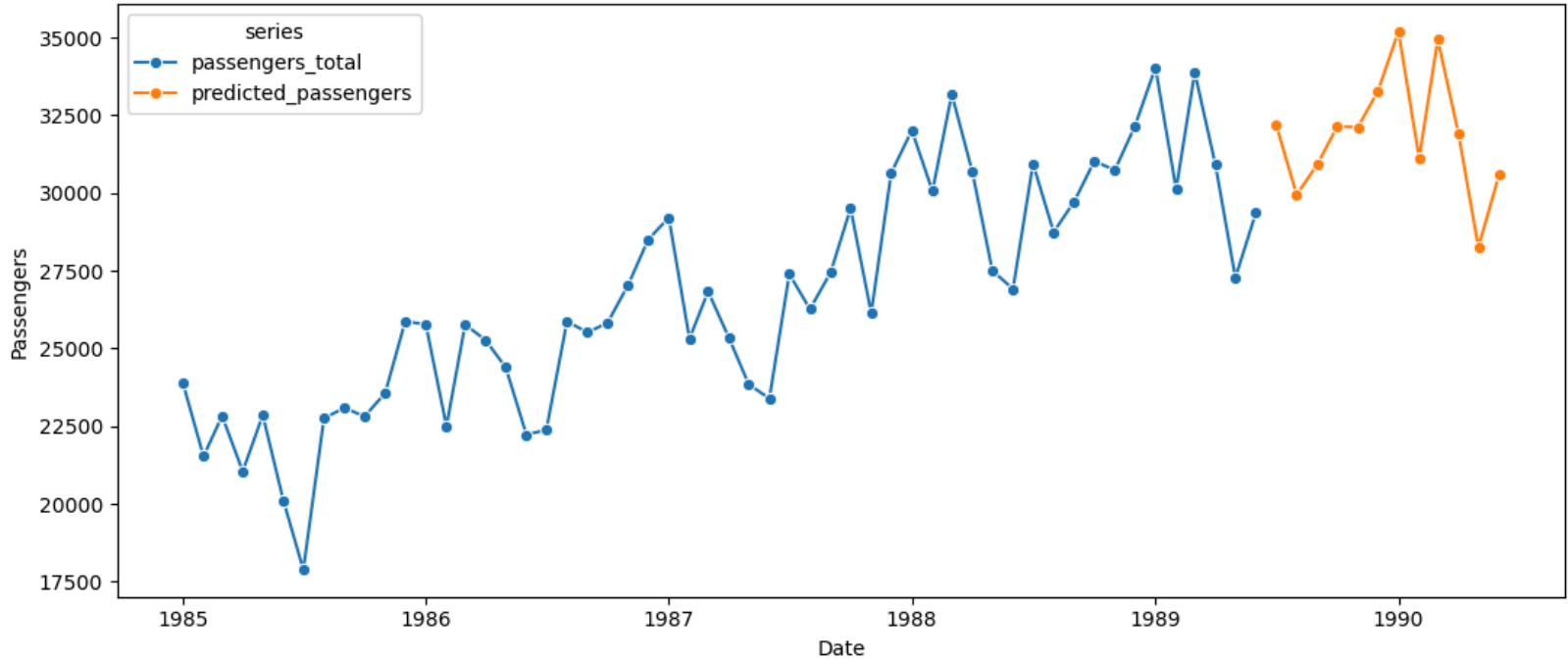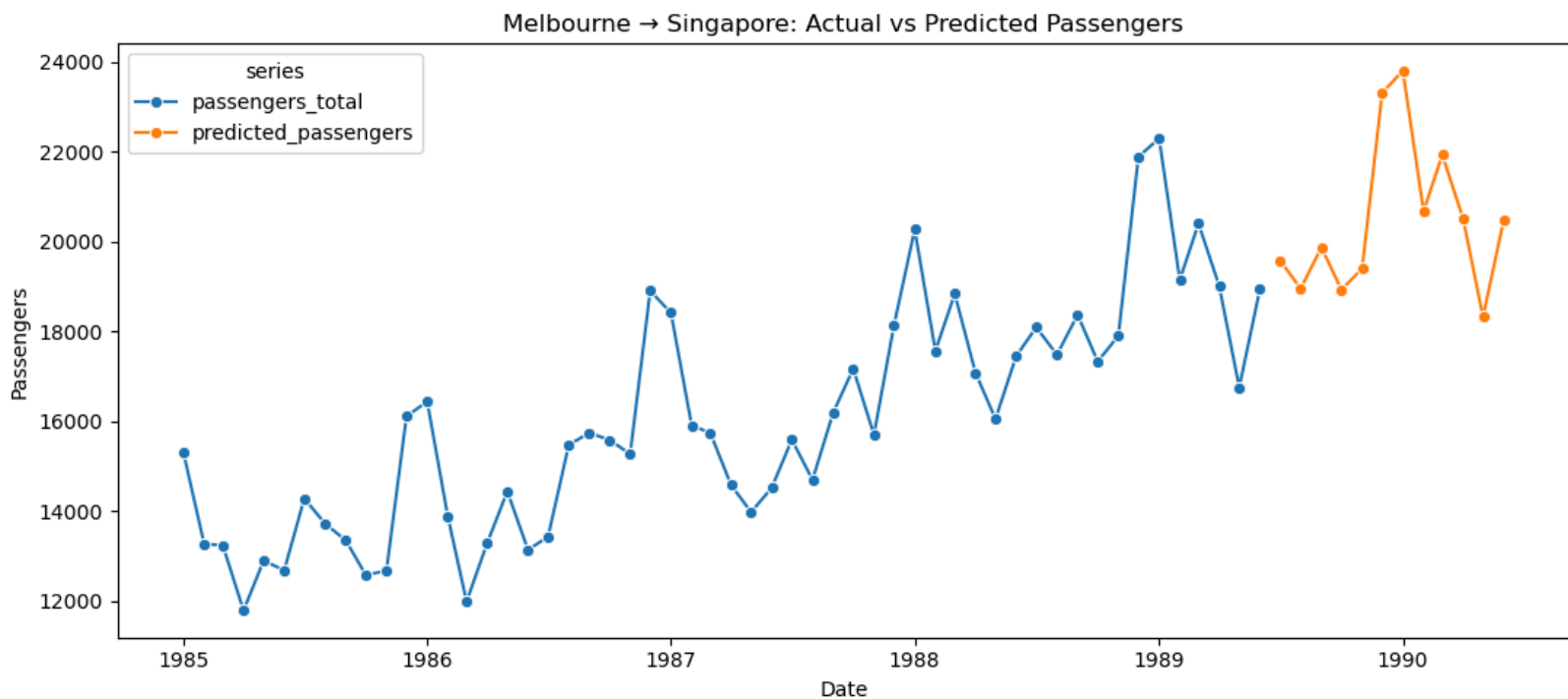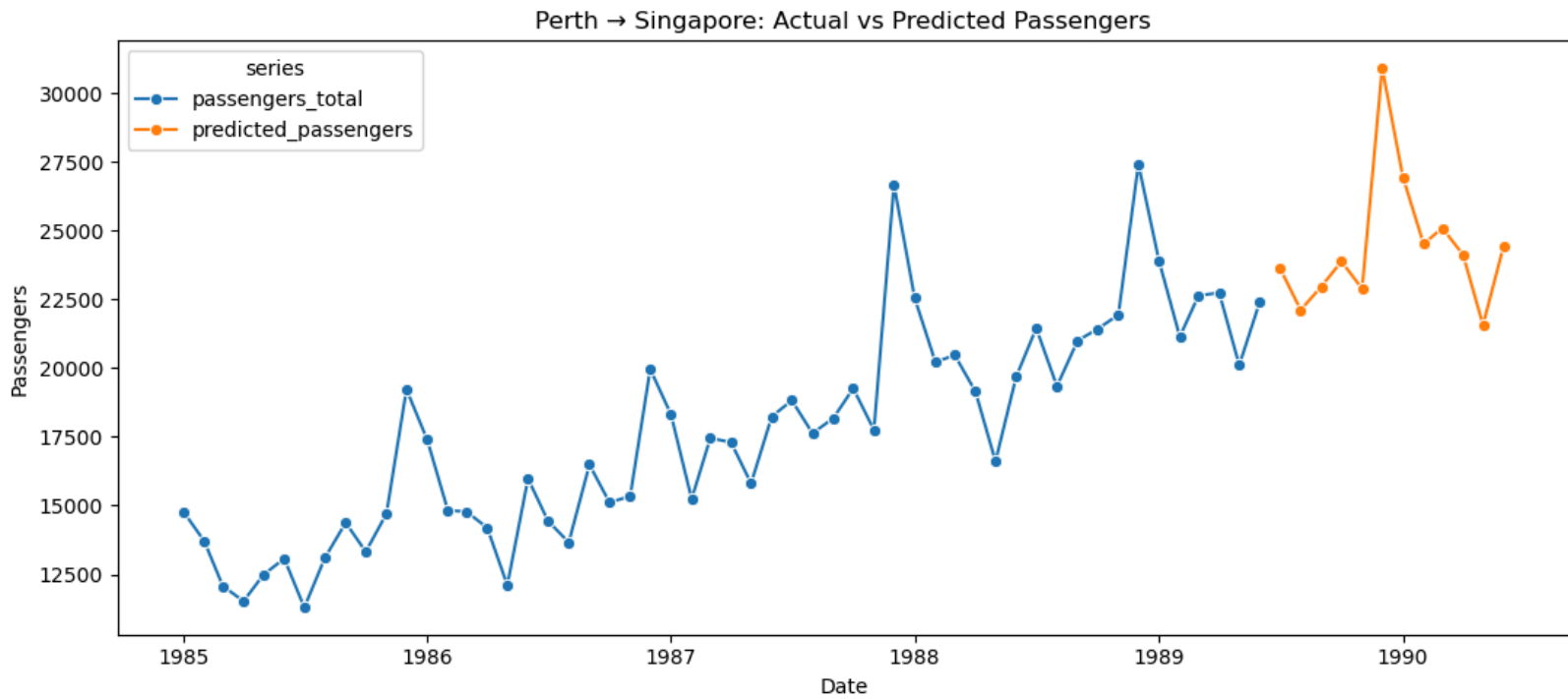
3. **Evaluate your model**

   a) Explain your model choices — why did you choose the elements you did
   Since the dataset was related to time series and each point update on a monthly basis, I
   knew that a model that accounts for seasonal change needed to be employed for this data.
   I knew linear regression wasn't going to work here since it doesn't account for volatile

peaks and valleys in seasonal data like this and only tries its best to fit the data based on the standard deviation, outliers, and residuals of the data. So I decided SARIMA was the best fit here, especially for city pairs with Singapore since the general trend each year is very stable.

The parameters I used in this model for the non-seasonal order was (1,1,1), where **p=1** (first parameter) was for one autoregressive term, meaning that the current value depends partly on the most recent month. For **d=1** (second parameter), since the time series plot had a general upward trend, I wanted to use first differencing to remove any long-term trend that could distort the stationarity of the series (since the mean is continually increasing). Finally, **q=1** (third parameter) accounted for one moving average term, which allowed the model to capture and correct patterns in the residual errors from prior time steps.

For the seasonal order, the explanations are very similar to non-seasonal order, and I used (1,1,1,12). Here, **P=1** (first parameter) included one seasonal autoregressive term, which lets the model use the same month of the prior year as a predictor. **D=1** (second parameter) applied seasonal differencing, which helps remove repeating yearly cycles and stabilizes the mean across seasons. **Q=1** (third parameter) introduced one seasonal moving average term, so the model could account for error terms from the same season in the past. Lastly, **s=12** (fourth parameter) was chosen because the dataset was monthly and clearly had annual seasonality, with consistent peaks during December/January and dips in February.

Since I was able to use the short term (p,d,q) and seasonal (P, D, Q, s) components, the SARIMA model was able to capture the month to month fluctuations and the larger yearly patterns observed in the past, making stable seasonal trends extremely accurate.

b) Evaluate the model's performance & report the accuracy of the model
Overall, the model did an excellent job in predicting the city pairs of Melbourne, Perth, and Sydney with Singapore. Their mean average percent error (MAPE) was between 2 to 4%, which means that the predicted monthly passenger counts were, on average, only slightly off from the actual values. This shows how accurate the model is when there's a consistent seasonal trend shown. I found this value by splitting the data by training the first four years of the data (1985-1988) and testing the first six months of 1989.

While the Singapore tests were successful, this model wasn't as accurate when predicting the passenger traffic for Sydney to Auckland as the relative maximums and minimums did not occur on the same months each year, and 1989 showed unusual volatility for this specific route that the model couldn't unfortunately account for. Overall, the model will perform well on time series trends where seasonality is stable and repeats across multiple years and might struggle slightly when there's not a consistent pattern each season.