Keshav Goel
Generate Data Challenge
08/21/2025

Task 1. Understanding the Data

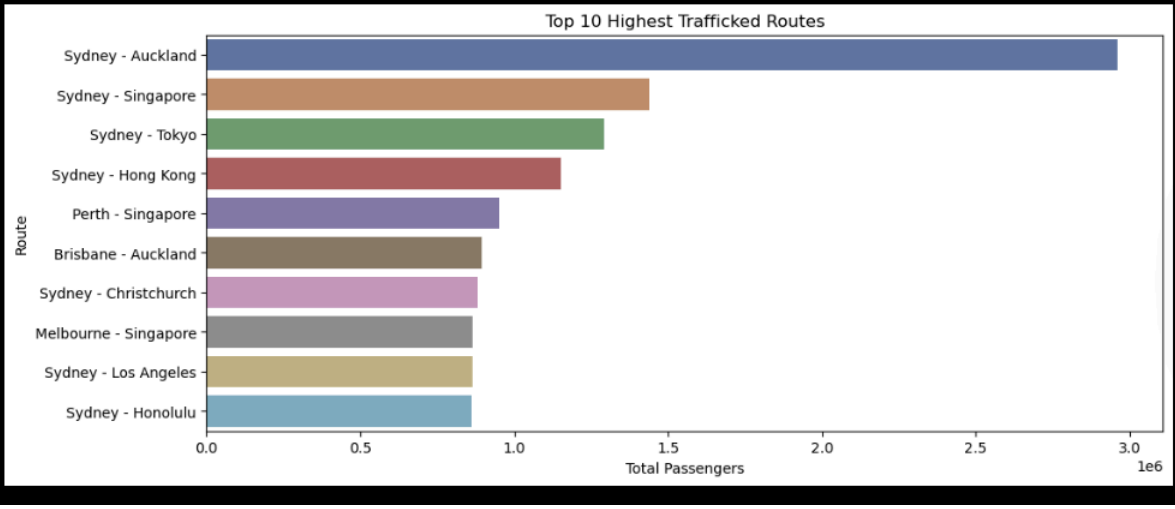
1a) Identify the most and least trafficked routes

Most

```
Top 10 Routes:
 route
Sydney - Auckland        2961212
Sydney - Singapore       1440018
Sydney - Tokyo           1292116
Sydney - Hong Kong       1151900
Perth - Singapore         952926
Brisbane - Auckland       893246
Sydney - Christchurch     882357
Melbourne - Singapore     865251
Sydney - Los Angeles      862964
Sydney - Honolulu         861814
Name: Passengers_Total, dtype: int64
/var/folders/bb/hsp4zv4n1j90qvy7v635c_m40000gn/T/ipykernel_90521/3967396550.py:11: F

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14

  sns.barplot(x=top_routes.values, y=top_routes.index, palette='deep')
```



Based on my exploration, the most trafficked route was from Sydney to Auckland. In general, the top routes consist of traveling from major city to major city. This is most likely because these are all highly popular travel destinations for tourists and businesses alike.

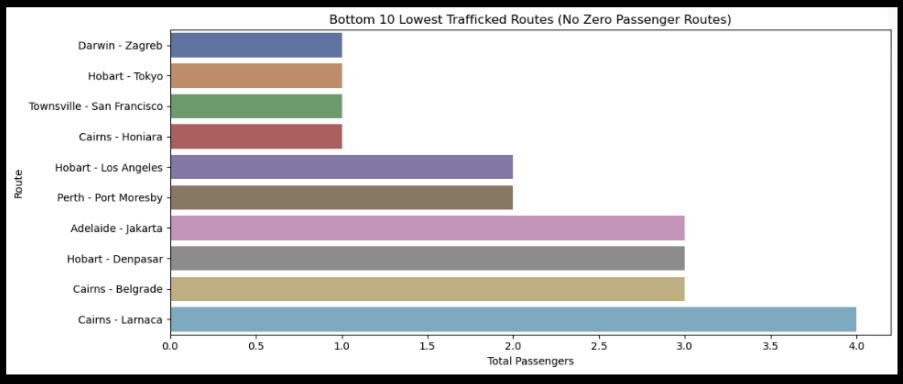
Least

```
Bottom 10 Routes with 0:
 route
Hobart - Los Angeles          2
Hobart - Tokyo                1
Cairns - Honiara              1
Townsville - San Francisco    1
Darwin - Zagreb               1
Melbourne - Denver            0
Perth - Bandar Seri Begawan   0
Brisbane - Colombo            0
Adelaide - Harare             0
Brisbane - Chicago            0
Name: Passengers_Total, dtype: int64

Bottom 10 Routes, no 0:
 route
Darwin - Zagreb               1
Hobart - Tokyo                1
Townsville - San Francisco    1
Cairns - Honiara              1
Hobart - Los Angeles          2
Perth - Port Moresby          2
Adelaide - Jakarta            3
Hobart - Denpasar             3
Cairns - Belgrade             3
Cairns - Larnaca              4
Name: Passengers_Total, dtype: int64
/var/folders/bb/hsp4zv4n1j90qvy7v635c_m40000gn/T/ipykernel_90521/1145131071.py:11: F

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14

  sns.barplot(x=bottom_routes_filtered.values, y=bottom_routes_filtered.index, palet
```



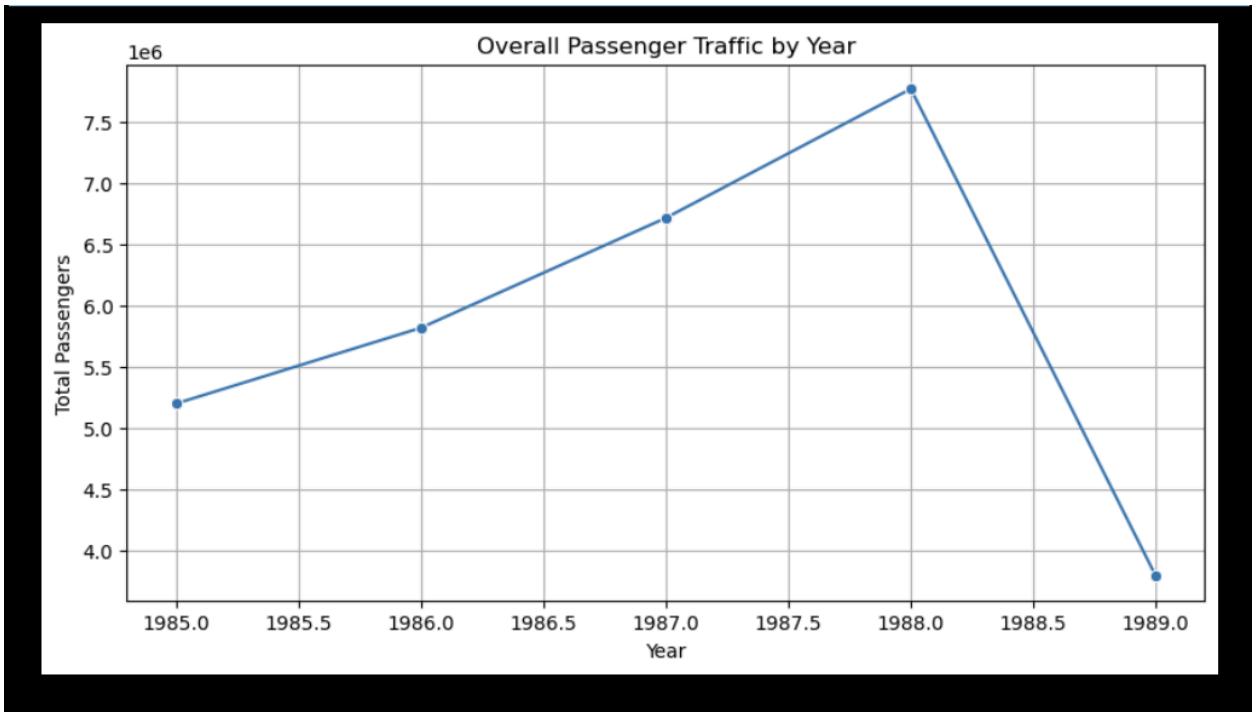
Based on my analysis, I decided to approach the lowest traffic routes in two separate ways. The first was analyzing routes that had zero passengers, and the second was analyzing routes including zero passenger ones. For the first facet, I found that Darwin-Zagreb, Hobart-Tokyo, Townsville - SF, and Cairns - Honiara all tied for last place with 1 passenger. While for the

second aspect, Melbourne-Denver, Perth-Bandar-Seri-Begawan, Brisbane-Colombo, Adelaide-Harare, and Brisbane-Chicago all tied for last place with 0 passengers. The possible reasons for this are varied, from less major cities to more export-focused than passenger-focused, to even a widespread airline strike that happened in Australia.

1b) Analyze Trends and/or geographical patterns

In this section, I will go over my visualizations, why I made them, and what they demonstrate.

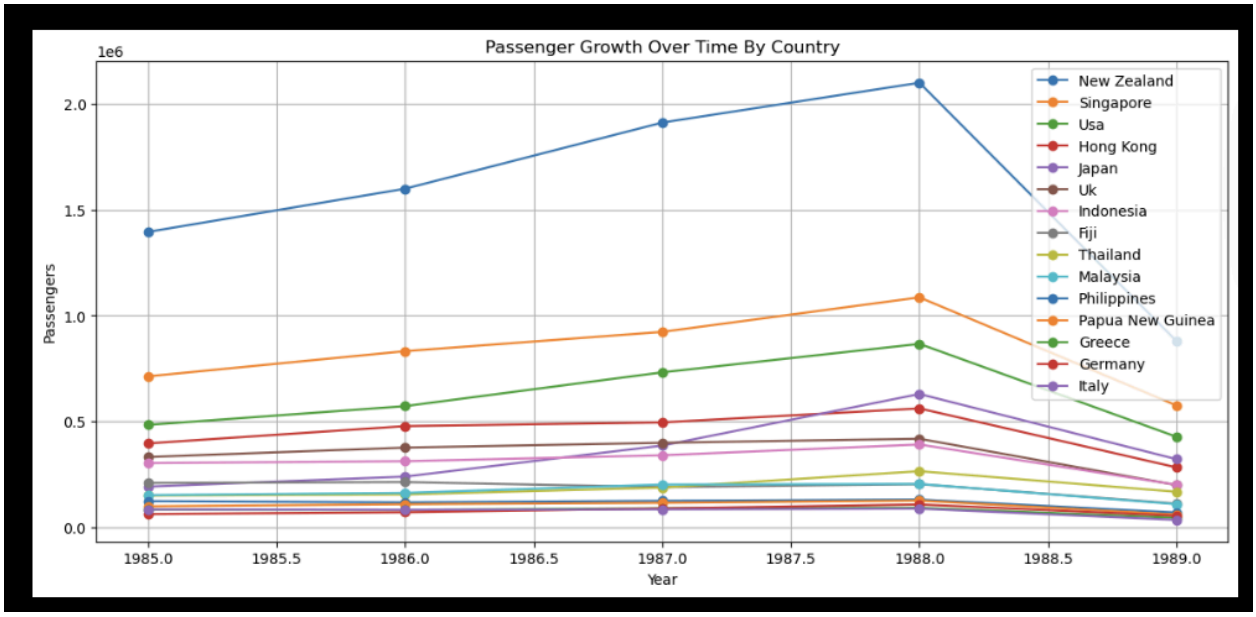
Visualization 1: Overall Passenger Traffic By Year



I created this line chart to get a high-level overview of AeroConnect's total passenger traffic across all routes and years. It sets the stage for more detailed route- or region-level analysis by first establishing an overall trend.

- Passenger traffic grew steadily from 1985 to 1988, showing an expansion in demand
- 1988 was the peak year with nearly 7.8 million passengers
- A sharp decline occurred in 1989, which was likely due to the airline strike
- Suggests that while growth exists, AeroConnect must be ready for volatile events like strikes

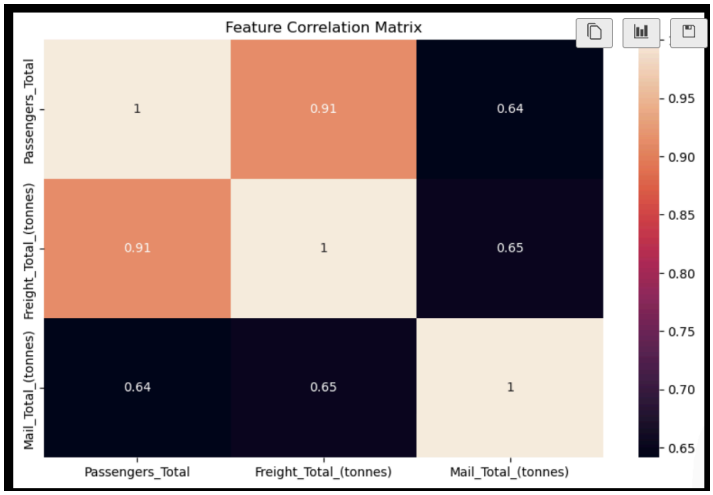
Visualization 2: Passenger Growth Over Time by Country



I created this line chart to break down passenger growth by country, allowing AeroConnect to see which international markets contributed most to traffic and how demand evolved across the globe.

- New Zealand dominates traffic volume, with consistent growth until 1988/1989, again likely due to strikes in Australia
- Singapore and the US show great growth as well, reflecting major regional hubs and long-haul markets
- Several other countries, like Hong Kong, show moderate and steady growth, showing a stable demand
- Every country dropped in 1989
- Shows AeroConnect's reliance on certain core markets and suggests opportunities for growth

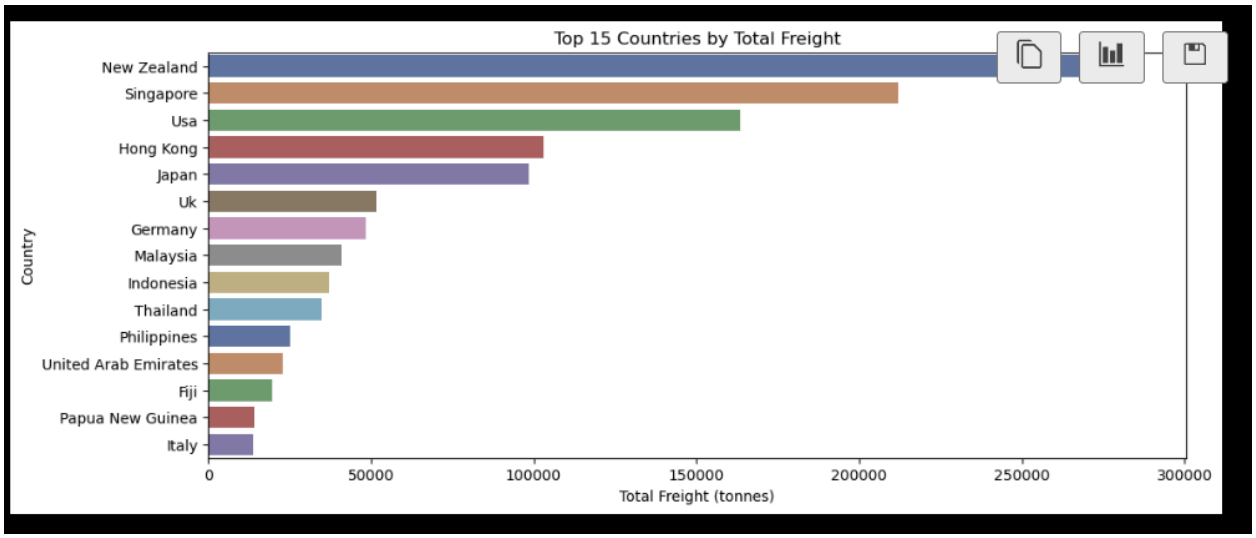
Visualization 3: Feature Correlation Matrix



I made this heatmap to examine how different key metrics are related to each other. This helps determine whether certain features move together and can provide insights into underlying patterns in airline operations.

- Passengers and freight are very strongly correlated (0.91), suggesting that busy passenger routes often also carry more cargo
- Mail traffic is moderately correlated with both, suggesting that while it could be connected, it's not tightly bound
- These highlight AeroConnect's strong passenger and freight operations, which are also closely intertwined. It could help reflect how resources are allocated.

Visualization 4: Top 15 Countries by Total Freight

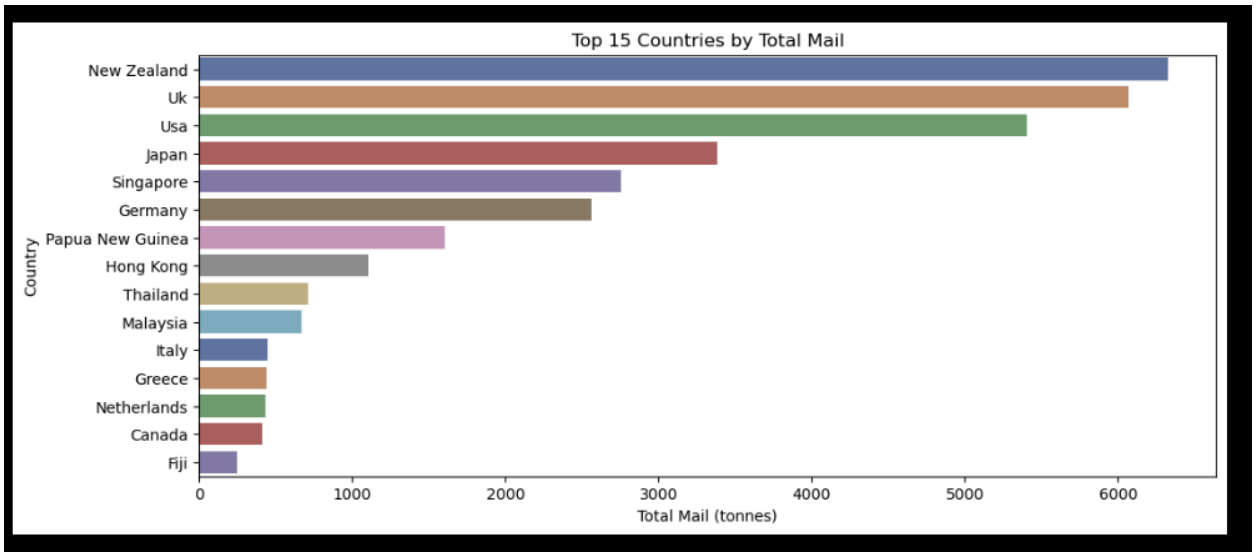


I created this horizontal bar chart to identify which international markets handle the most freight volume. By focusing on freight separately from passengers, this helps AeroConnect assess the importance of cargo operations and pinpoint key trade partners.

- New Zealand leads in freight volume, followed closely by Singapore and the USA, reinforcing their positions as major trade corridors.
- Other strong freight partners include Hong Kong, Japan, and the UK, all of which are global commerce hubs.
- Mid-tier contributors such as Malaysia, Indonesia, and Thailand indicate growing regional demand for cargo services.
- Smaller markets like Papua New Guinea and Fiji handle significantly less freight, suggesting these are niche routes.
- The chart highlights that AeroConnect's freight business is concentrated in a few high-volume countries, with opportunities for diversification in emerging Asian markets.

-

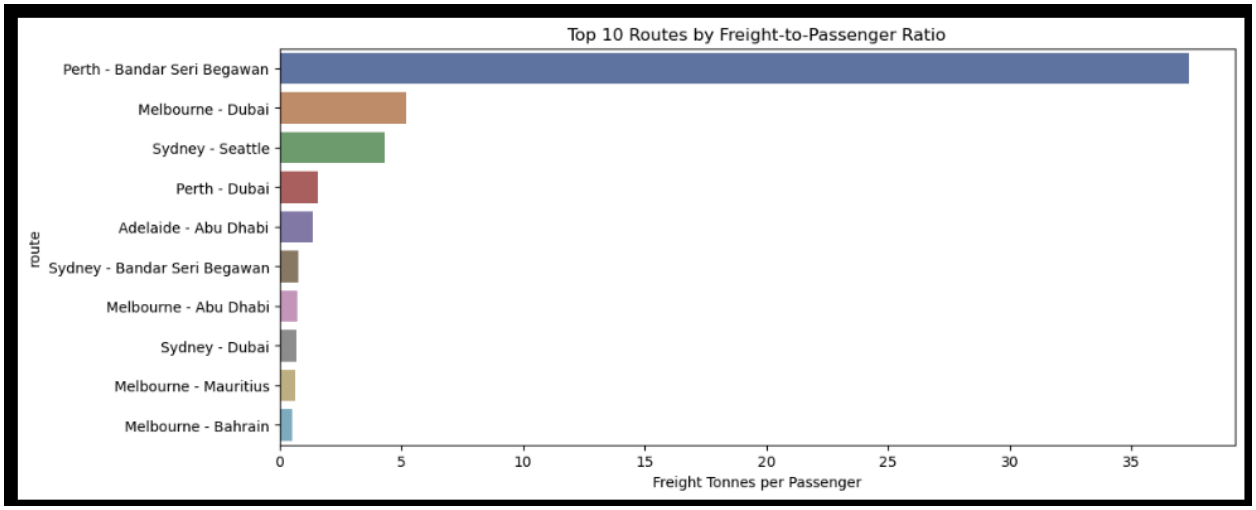
Visualization 5: Top 15 Countries by Total Mail



I created this chart to analyze AeroConnect's mail traffic by country, which helps highlight postal and logistics connections that may not align directly with passenger or freight flows.

- New Zealand once again dominates, showing strong postal ties alongside passenger and freight demand
  The UK and the USA are also major mail destinations, reflecting significant international correspondence and business connections
- Japan, Singapore, and Germany rank highly, consistent with their roles as global trade and communication hubs
- Interestingly, Papua New Guinea appears as a relatively strong mail partner despite being a smaller passenger and freight market, suggesting mail-specific demand.
- Lower-volume contributors include European markets and regional neighbors
- This chart shows that AeroConnect's mail business has both overlapping and distinct patterns from freight and passenger flows, with certain routes (e.g., PNG) driven primarily by postal needs

Visualization 6: Top 10 Routes by Freight-to-Passenger Ratio

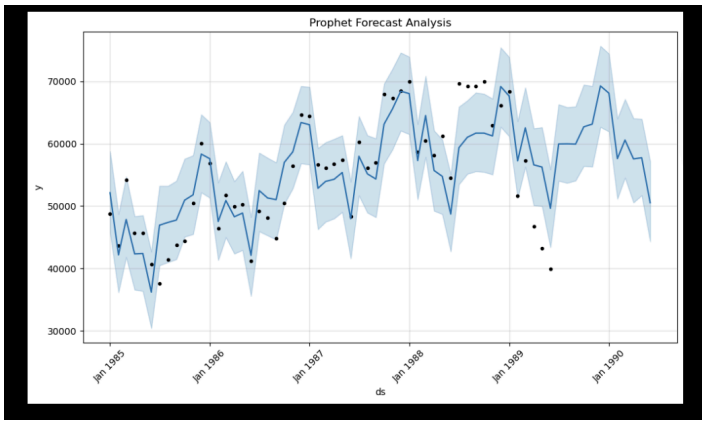


I built this bar chart to identify routes where freight demand significantly outweighs passenger demand. This helps AeroConnect understand which routes are cargo-heavy and may require different resource allocation strategies.

- The Perth–Bandar Seri Begawan route is an extreme outlier, carrying far more freight relative to passengers than any other route
- Other routes, such as Melbourne–Dubai and Sydney–Seattle, also have high freight-to-passenger ratios, making them strong candidates for cargo-focused services
- The rest of the routes show moderate ratios, but still skew toward freight compared to the majority of AeroConnect's network
- This suggests AeroConnect could consider dedicated freighter aircraft or optimized cargo services for certain international routes, rather than focusing solely on passenger flights

Task 2. Building the Model

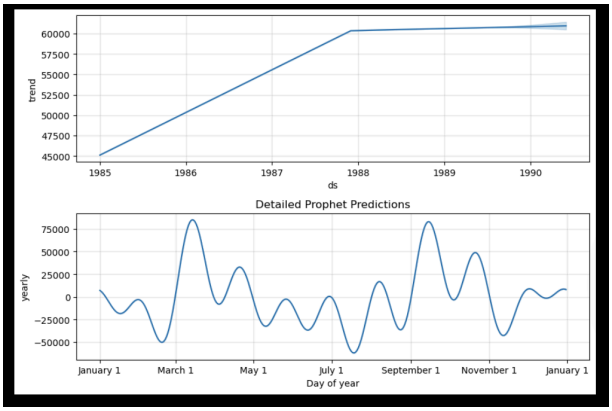
2a) Your model should predict passenger traffic for the next 6-12 months on at least 1 city pair



This graph shows the predictions that Prophet made. The solid blue line represents the predictions, the black dots represent the actual, and the shaded area represents the uncertainty interval. From the graph, the spikes and drops seem consistent with the rest of the prediction.

```
12 month forecast:
           ds          yhat    yhat_lower    yhat_upper
54 1989-07-01  59969.993910  54048.873181  66311.915007
55 1989-08-01  59993.330106  53714.349910  65865.395123
56 1989-09-01  59950.005006  54095.453263  65972.865232
57 1989-10-01  62723.166182  56423.648748  69474.783240
58 1989-11-01  63156.531603  56323.169619  69253.172335
59 1989-12-01  69281.568040  62682.835507  75687.565242
60 1990-01-01  68123.374753  61963.695099  74449.907717
61 1990-02-01  57616.887483  51181.319888  64039.148674
62 1990-03-01  60602.485929  54547.848083  67143.820195
63 1990-04-01  57570.378414  50567.439720  64069.552695
64 1990-05-01  57758.644120  51721.597380  63975.499178
65 1990-06-01  50542.184284  44343.558594  57167.087113
```

12-month prediction values from Prophet. Yhat represents Prophet's predictions from July 1989 to June 1990. Yhat_lower and Yhat_upper provide a confidence interval and range of possible outcomes.



These graphs represent a more detailed view of Prophet. The top plot shows the long-term trend with passenger traffic slowly growing from 1985-1988 and then sort of flattening out from 1989 onwards. The Bottom plot represents the Yearly seasonality by showing the spikes and dips throughout the year. Peaks are the major travel season, while troughs are the off-peak months

Task 3. Evaluate Your Model

3a) Explain your model choices – why did you choose the elements you did

I chose Meta Prophet as my forecasting model because I wanted to explore a tool outside of the usual scikit-learn regressors. During my co-op, one of my coworkers introduced me to Prophet and told me about its ease of usability for time-series analysis, since he was using it to predict yearly pharmaceutical sales at major pharmacies like CVS. This made Prophet a natural choice for this assignment, given its strength in handling seasonal and trend-based patterns, as well as its ease of customization.

Also, unlike traditional ML workflows, I did not need to use a manual train-test split because Prophet has its own built-in cross-validation framework for time-series data. This is important

because normal cross-validation would break the temporal order of the data, while Prophet's methodology respects the sequential nature of the data.

For this analysis, I applied Prophet to the Sydney-Auckland route, the top passenger route in the dataset, as it provided a meaningful case study as well as a lot of data to train our model on. Since Prophet requires the target variable y and the date ds, I had to reformat the data into the expected structure inside a respective Prophet Dataframe. When initializing the model, I enabled yearly seasonality, while disabling daily and weekly, since travel demand is usually seasonal at the yearly level (holidays, break, business) rather than based on weekly or daily factors. Furthermore, Prophet's inclusion of confidence intervals around forecasts adds a layer of interpretability and allows AeroConnect to plan with a better understanding of uncertainty.

3b) Evaluate the model's performance & report the accuracy of the model

To evaluate the model, I compared Prophet's predicted passenger totals (yhat) with the actual observed values (y) using common metrics:

R2 = 0.717

- This means that the model explains about 72% of the data. While it's not perfect, it is a strong fit given the volatility of airline data and how far in the future we were trying to predict.

MSE = 24,071,540

- This measures the average squared difference between predicted and actual passenger counts. An error of this scale means that some month predictions were off by tens of thousands of passengers.

MAE = 3892.243

- This measures that Prophet's predictions, on average, were off by about 3892 passengers per month. Considering that the monthly traffic on the Sydney-Auckland route ranges from 50,000 - 70,000, this is an error of about 5-7%, which is quite good for operational forecasting.

Overall, Prophet provided a reasonably accurate model for forecasting air traffic while also maintaining a pretty simple ease of use. It effectively captured both the long-term growth trend and the seasonal spikes, while maintaining relatively low prediction errors compared to the overall passenger volumes. This makes it a reliable tool for AeroConnect to use in short-term planning, though its limitations should be noted for routes with more volatility or smaller datasets.

Task 4. Provide Recommendations

4a) Which Routes should AeroConnect invest more in or scale back from?

Based on the overall analysis, AeroConnect should prioritize investment in routes that consistently show high and growing passenger demand and/or strong alignment across passenger, freight, and mail traffic.

- Invest more in:
    - Sydney-Auckland: The top-performing route by passenger volume, and has a reliable growth market
    - Melbourne-Sydney and Sydney-Singapore: High-traffic routes with strong freight components
    - Focus more on markets like the USA and Japan, which demonstrate steady and consistent growth over time

- Scale back from:
    - Routes with consistently low demand, such as smaller regional connections like Fiji, Papua New Guinea. The routes don't really justify large aircraft allocations and can be better served with less frequent routes or smaller planes.
    - Cargo-heavy but passenger-light routes like Perth-Bandar Seri Begawan might be better suited for a plane that focuses on cargo rather than passenger services.

4b) How can AeroConnect use this model going forward?

The Prophet model can serve as a valuable forecasting tool to guide both short- and long-term operations/strategy.

- Operations: Prophet can help determine overall planning, like flight volume, aircraft size, and more
- Revenue: It can help anticipate demand spikes and drops,so that AeroConnect can adjust ticket pricing and promotional strategy.
- Cargo: Routes with high freight-to-passenger ratios can be earmarked for specialized freighter services or hybrid passenger-cargo operations
- Market: If we continuously retrain Prophet with new data, AeroConnect can track emerging demand in secondary markets like Malaysia and decide when to increase capacity
- Risk: With confidence intervals, AeroConnect can view an early-warning system for uncertainty. Allowing managers to plan better when predictions suggest a possible downturn