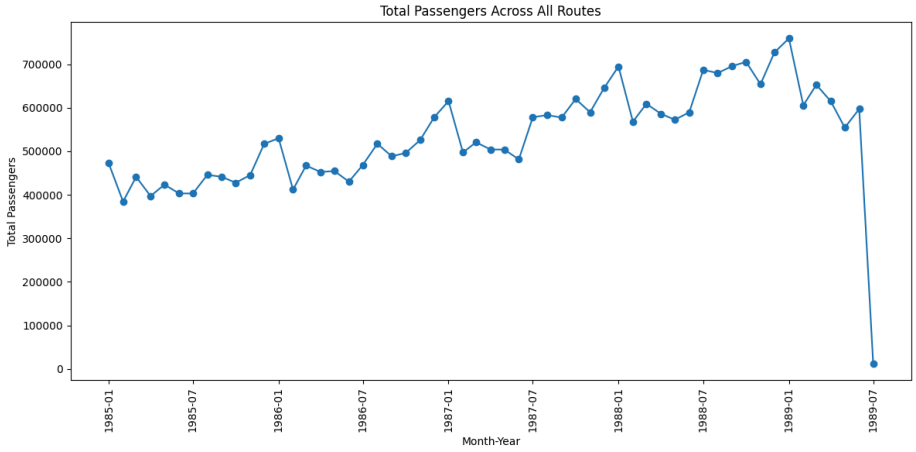
Generate Data Technical Challenge - Hannah Chacko

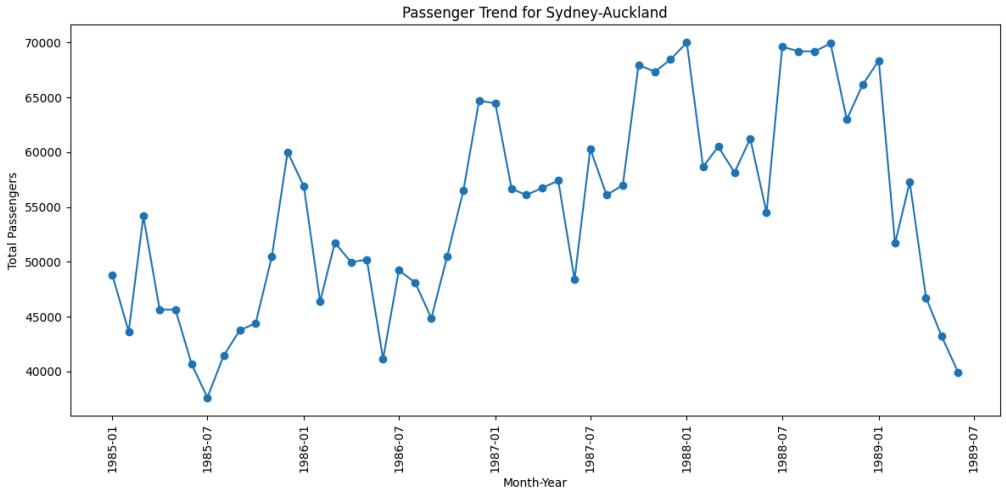
1. Understanding the Data
    a. Identify the Most and Least Trafficked Routes

    The most active route was Sydney-Auckland with 2,961,212 passengers total. The least active routes, with only 1 passenger each, were Darwin-Zagreb, Hobart-Tokyo, Townsville-San Francisco, and Cairns-Honiara.

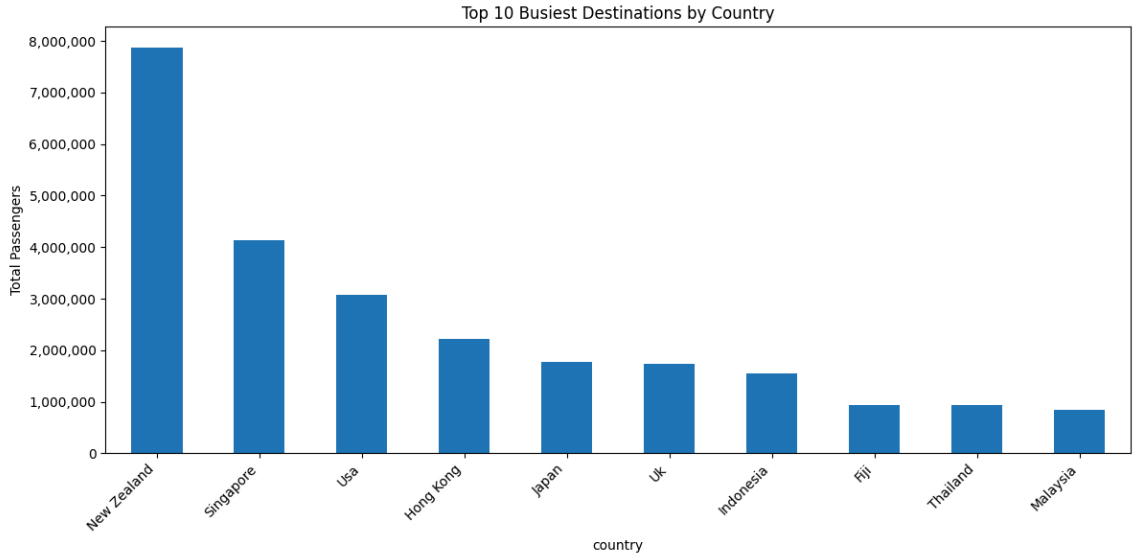
```
Top 10 most active routes:
                   route  passengers_total
228      Sydney-Auckland           2961212
280     Sydney-Singapore           1440018
281         Sydney-Tokyo           1292116
249     Sydney-Hong Kong           1151900
218      Perth-Singapore            952926
37     Brisbane-Auckland            893246
237  Sydney-Christchurch            882357
179  Melbourne-Singapore            865251
259   Sydney-Los Angeles            862964
251      Sydney-Honolulu            861814


Top 10 least active routes: (excluding trips with 0 passengers)
                       route  passengers_total
125            Darwin-Zagreb                 1
134             Hobart-Tokyo                 1
303  Townsville-San Francisco                1
87             Cairns-Honiara                1
215        Perth-Port Moresby                2
130        Hobart-Los Angeles                2
128           Hobart-Denpasar                3
82            Cairns-Belgrade                3
16           Adelaide-Jakarta                3
90             Cairns-Larnaca                4
```

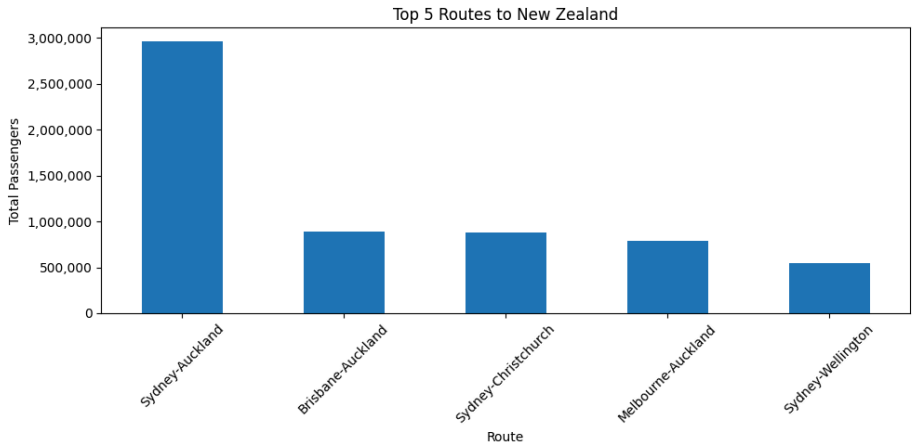
    b. Analyze trends and/or geographical patterns



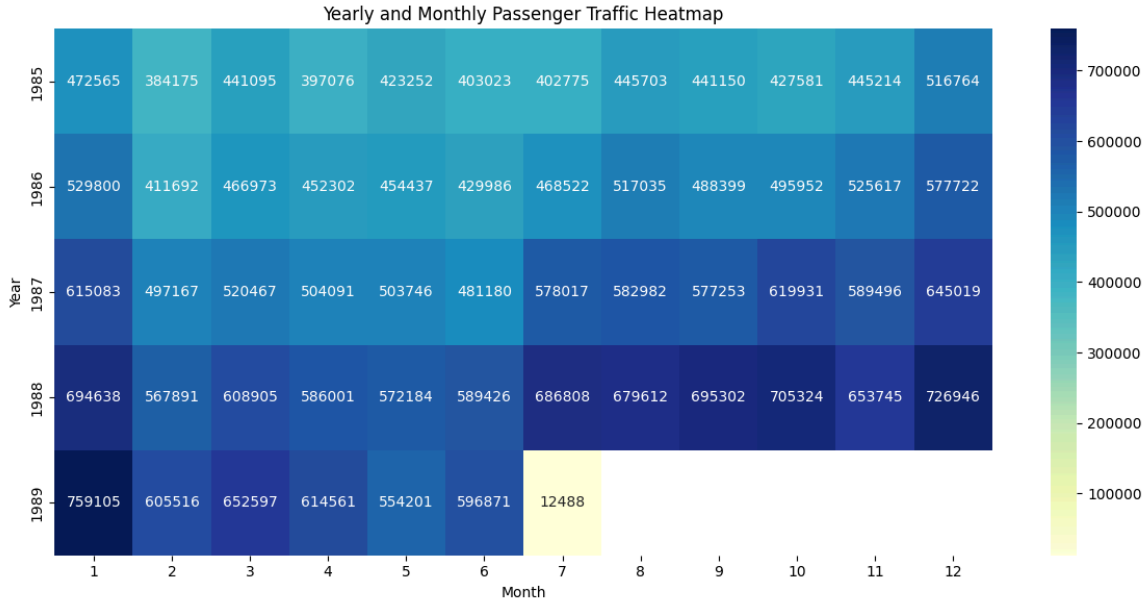


This visualization of total passengers across all routes shows an overall picture of the seasonal and yearly trends. This plot is also customizable for certain routes which is useful for identifying patterns for specific routes. Overall, this is useful for identifying periods of high or low demand which can help AeroConnect schedule their trips and allocate the proper resources needed.

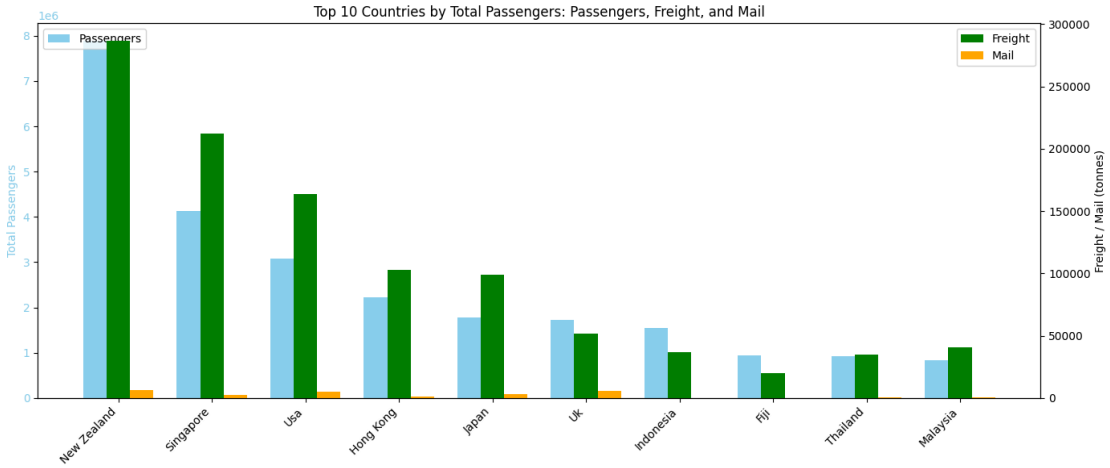


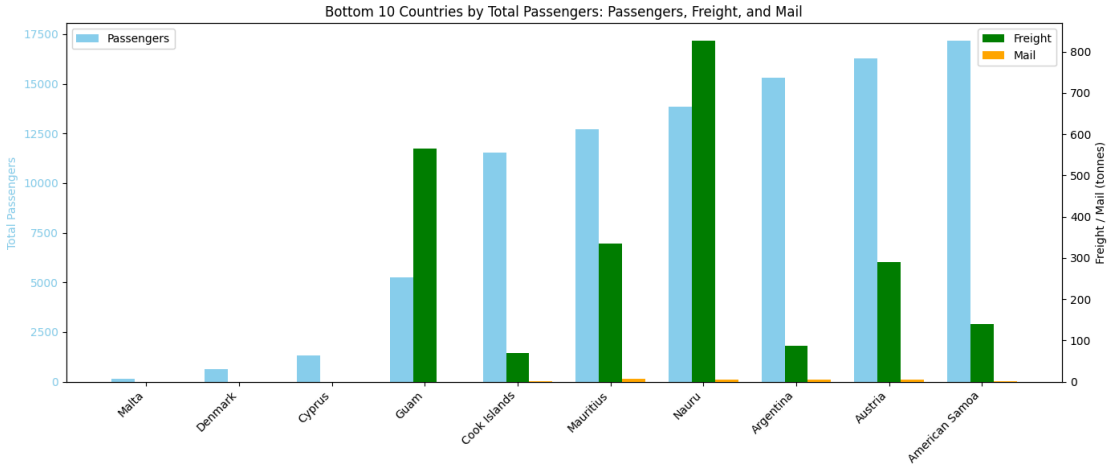


The bar chart of top countries help visualize the busiest countries served, helping plan strategic operations to ensure maximum profit. Then, delving into the most popular routes into the top countries provides a clean view of where the demand is necessary within a country, enabling targeted improvements in service.



This heatmap effectively visualizes the seasonal variations across multiple years, allowing for analysis to detect holiday spikes or off-peak months which supports long term forecasting and trend analysis.



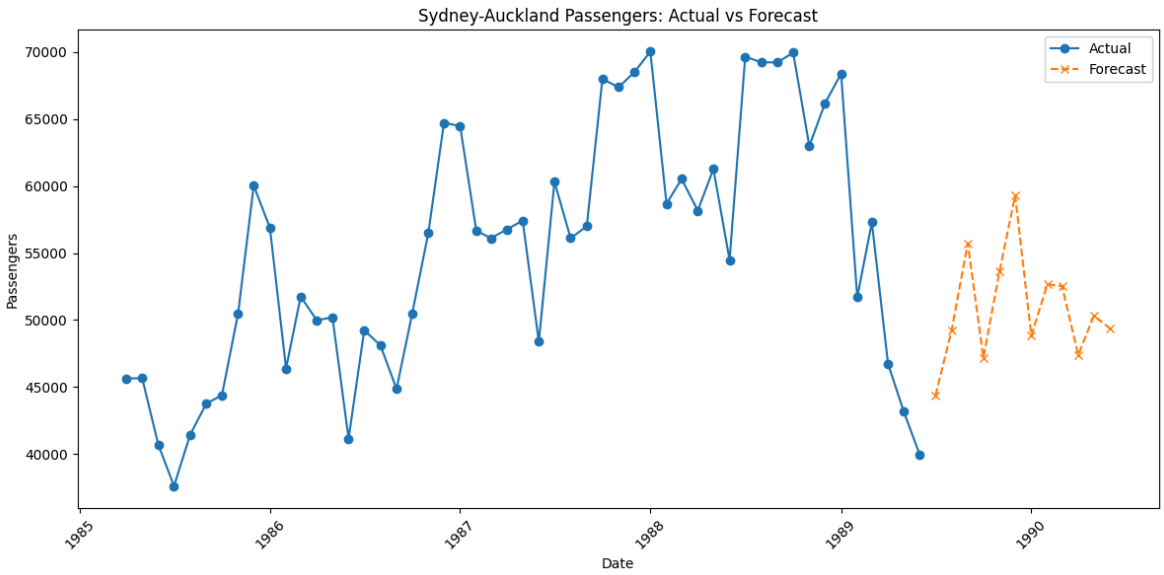


This grouped bar chart compares total passengers, freight, and mail across different countries, highlighting the multiple revenue streams AeroConnect operates in. It shows that the airline's activity is not limited to passenger transport but also includes significant freight and mail operations. By analyzing these trends, AeroConnect can identify opportunities to optimize cargo and mail allocation, adjust service levels, and strategically target routes that maximize overall profitability.

2. Model

a.

```
Random Forest Performance for Sydney-Auckland:
Best Parameters: {'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300}
Mean Squared Error (MSE): 8,022,452.99
Root Mean Squared Error (RMSE): 2,832.39 passengers
Mean Absolute Error (MAE): 2,343.48 passengers
Mean Absolute Percentage Error (MAPE): 4.39%
Mean monthly passengers: 55,188
RMSE as % of mean passengers: 5.13%
R^2 Score: 0.9077
```



b.

3. Explain Your Model
    a. I went with a Random Forest model to predict Sydney-Auckland passenger traffic because it handles patterns that aren't perfectly straight lines and can deal with ups and downs in the data. I used lag features so the model can learn from the last few months, and included month and year to capture seasonal trends and yearly changes. I also used a time series split so the model only trains on past data when predicting future months, which keeps it realistic, and did hyperparameter tuning to find the best tree depth, number of trees, and splitting rules so it doesn't overfit or underfit the data.
    b. The model performed really well for this route, which was the most active with nearly 3 million passengers in total. The RMSE was 2,832 passengers per month, and the mean absolute percentage error was just 4.39%, meaning the predictions are very close to the actual monthly traffic. The R^2 score was 0.9077, showing that the model explains about 91% of the variation in passenger numbers. Compared to the mean monthly passengers of 55,188, the RMSE is only around 5% of that value, so the forecast is accurate enough to support planning decisions. This means AeroConnect can rely on the model to predict upcoming passenger traffic, helping them optimize routes, adjust capacity, and make informed investment choices.
4. Provide Recommendations
    a. Route Investment Recommendations:
    Looking at the numbers, Sydney-Auckland is by far the busiest route with almost 3 million passengers. Since traffic here is high and predictable, it makes sense for AeroConnect to invest more by adding more flights or using bigger planes to carry more passengers. Conversely, routes like Darwin-Zagreb or Hobart-Tokyo barely have any passengers so it'd be smarter to scale back on these, either by reducing flights or reallocating planes to busier routes.

    b. How AeroConnect Can Use This Model Going Forward:

This Random Forest model is super useful for planning and decision-making. AeroConnect can use it to forecast monthly passenger numbers for different city-pairs, which helps figure out flight schedules, staffing, and plane allocation. And since the model can be retrained with new data, it’ll keep getting better over time which will help AeroConnect make smarter, data-driven choices to maximize revenue and resources.