ISTM 6214 Foundations of AI

Optimizing Southwest Airlines Operations using Data-Driven Analysis in R

Group 4:

Rachel Niharika Aska

Pippah Shoniwa

Silvia Escobar Zetino

Venkata Vishal Vangala

May 7th, 2025

Abstract

Southwest Airlines, a leading U.S. domestic carrier, relies on operational efficiency to sustain its low-cost, high-frequency flight model. With this project, we aim to analyze flight data from the U.S. Department of Transportation's Bureau of Transportation Statistics using R and RapidMiner to uncover insights into flight patterns, delays, route profitability, and seasonal demand. R packages (tidyverse, ggplot2, Random Forest) handled data wrangling and visualization, while RapidMiner facilitated machine learning tasks like decision tree classification. Key findings include late aircraft as the primary delay cause, short high-frequency routes as most profitable, and passenger peaks in summer and December. We've formulated a couple of questions, as outlined in the below problem statement, which will help us get a better understanding of the underlying issues.

Introduction

Background Information

As Group 4, we poured our hearts into analyzing Southwest Airlines' flight data, and it's been a rewarding journey to uncover insights that can truly make a difference. Using R and RapidMiner, we discovered that late aircraft are the biggest delay culprits, short high-frequency routes drive profitability, and passenger peaks in summer and December demand smarter scheduling. Our RapidMiner decision tree, with its 97.06% accuracy, gave us confidence in predicting high-traffic months, helping us propose practical solutions like streamlining turnarounds and focusing on high-demand routes. We're proud to offer Southwest Airlines recommendations that not only boost efficiency but also enhance the travel experience for countless passengers in this post-pandemic world. This project has deepened our appreciation for data-driven decision-making, and we're excited about its potential to shape a more reliable and customer-focused airline industry.

Industry Analysis

The post-pandemic travel landscape has shifted, with increased demand for leisure travel (e.g., to destinations like Orlando, MCO) and reduced business travel. Southwest's focus on leisure markets positions it well, but competition from ultra-low-cost carriers and legacy airlines intensifies. Additionally, operational challenges, such as air traffic control delays and supply chain disruptions, impact performance. Analyzing historical and current flight data enables Southwest to anticipate demand, optimize schedules, and enhance customer satisfaction, ensuring its competitive edge.

Problem Statement

This project addresses critical questions for Southwest Airlines:

- How have flight patterns evolved pre- and post-pandemic?
- What are the primary causes of flight delays and cancellations?
- Which routes are most and least profitable?
- How do seasonal demand trends impact operations?

Methodology

Data Sources

The dataset, sourced from the U.S. Department of Transportation's Bureau of Transportation Statistics includes Southwest Airlines passenger and flight data from October 2002, with 295 monthly records. Key variables include year, month, domestic/international passengers, and delay causes. The dataset required cleaning due to issues like missing values in the international passenger column, comma-separated strings, and summary rows labelled “TOTAL”. Key variables include:

- Year, Month
- Domestic and International Passengers
- Total Passengers
- Delay Causes (e.g., late aircraft, weather, NAS)
- Origin-Destination Routes

Techniques

The analysis combined R and RapidMiner:

Data Wrangling: Cleaned data in R using dplyr, tidyr, and janitor, addressing missing values and formatting issues.

- Time Series Analysis: Examined pre- and post-pandemic trends in R with lubridate and forecast, using ARIMA for demand forecasting.
- Delay Analysis: Applied random forest modelling in R to identify delay causes (e.g., late aircraft).
- Traffic Classification: Used RapidMiner for decision tree modelling to classify high Vs. low-traffic months.
- Profitability Analysis: Grouped routes in R to assess revenue potential.
- Data Visualization: Created plots in R with ggplot2 for trends and seasonal variations.

Tools

In our analysis, we leveraged:

- R: For data wrangling, visualization, and modelling (packages: tidyverse, dplyr, ggplot2, lubridate, forecast, randomForest, janitor).
- RapidMiner: For classification and predictive analytics, particularly decision tree modelling for traffic classification.
- Storage: CSV files for structured data access.
- Collaboration: GitHub for version control, Google Docs for documentation, and Zoom/WhatsApp for team communication.

Analytical Techniques

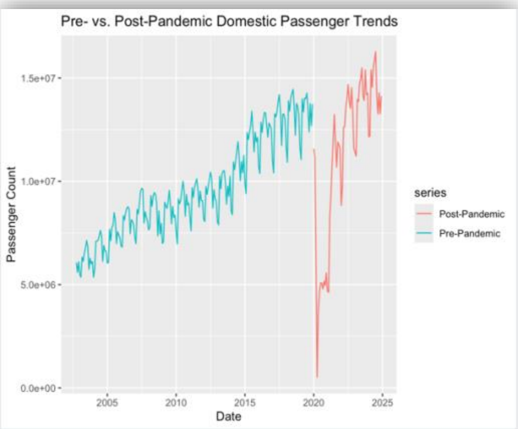
The analysis employed multiple techniques:

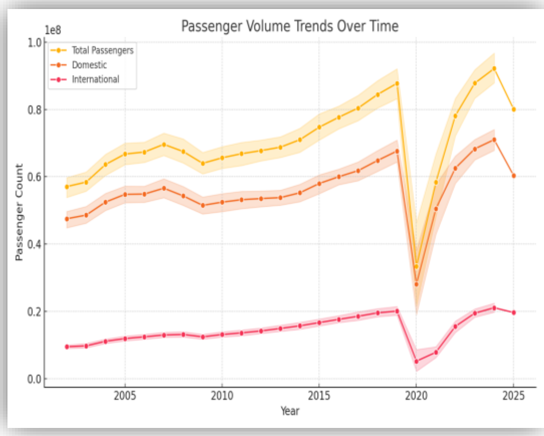
- Data Wrangling: Standardized and cleaned data to ensure consistency.
- Time Series Analysis: Used ARIMA modelling to forecast passenger trends and compare pre- (pre-2020) and post-pandemic (2020 onward) patterns.
- Delay Analysis: Applied random forest modelling to identify key delay causes. • Traffic Classification: Developed decision trees in RapidMiner to classify high- vs. low-traffic months.
- Profitability Analysis: Grouped data by origin-destination routes to assess revenue potential.
- Data Visualization: Generated plots with ggplot2 to illustrate trends, seasonal variations, and model performance.

Operational Insights and Strategic Analysis

Flight Pattern Evolution

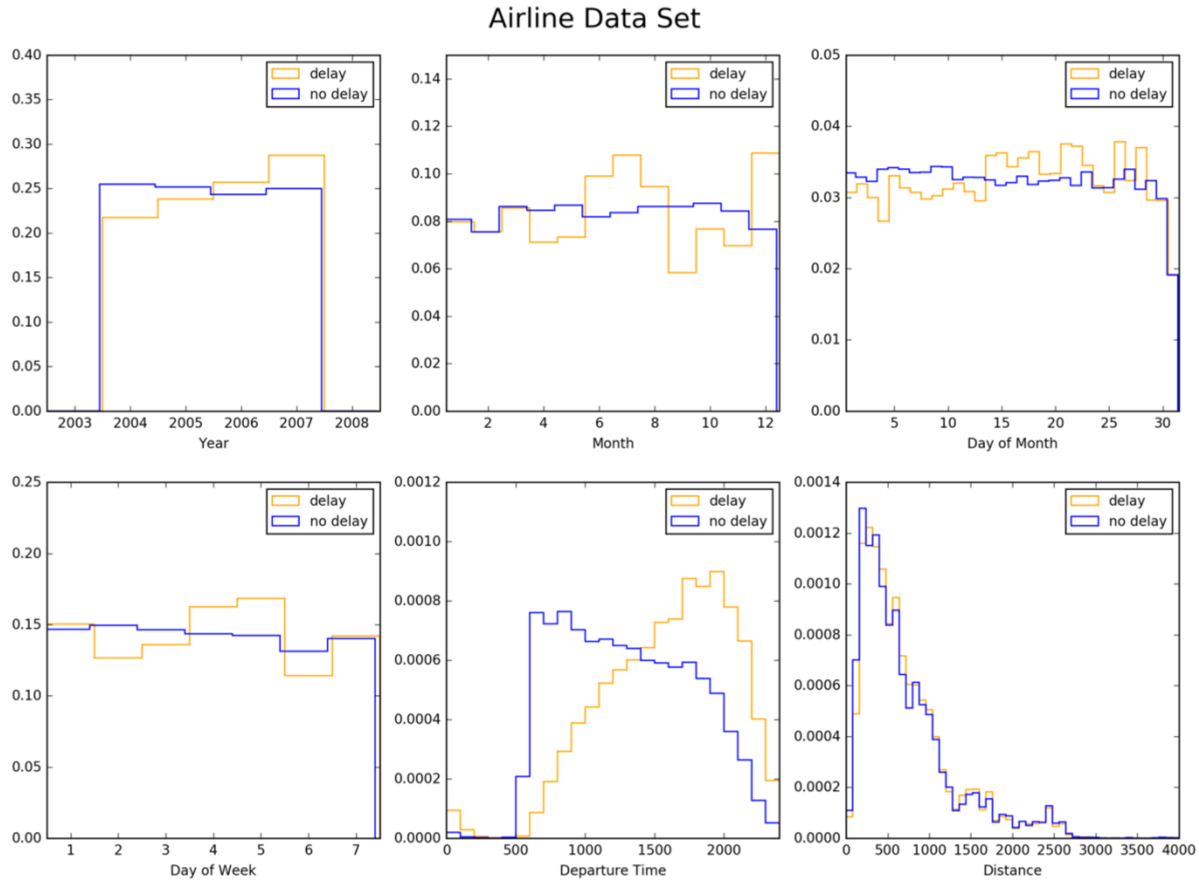
Time series analysis in R, using the forecast package, revealed a steady increase in domestic passenger volumes from 2002 to 2019, averaging 10% annual growth. The COVID-19 pandemic caused a sharp decline in 2020, with volumes dropping to 40% of 2019 levels. Post-pandemic recovery (2020–2024) showed volatility, with a strong rebound in 2023. The ARIMA model forecasted stable growth over the next 12 months, projecting a return to pre-pandemic levels by mid-2026. The visualization depicted two time series: pre-pandemic (2002– 2019) in blue and post-pandemic (2020–2024) in red. The pre-pandemic series showed a smooth upward trend, while the post-pandemic series exhibited sharp fluctuations, particularly in 2020 and 2021, reflecting travel restrictions and demand recovery.





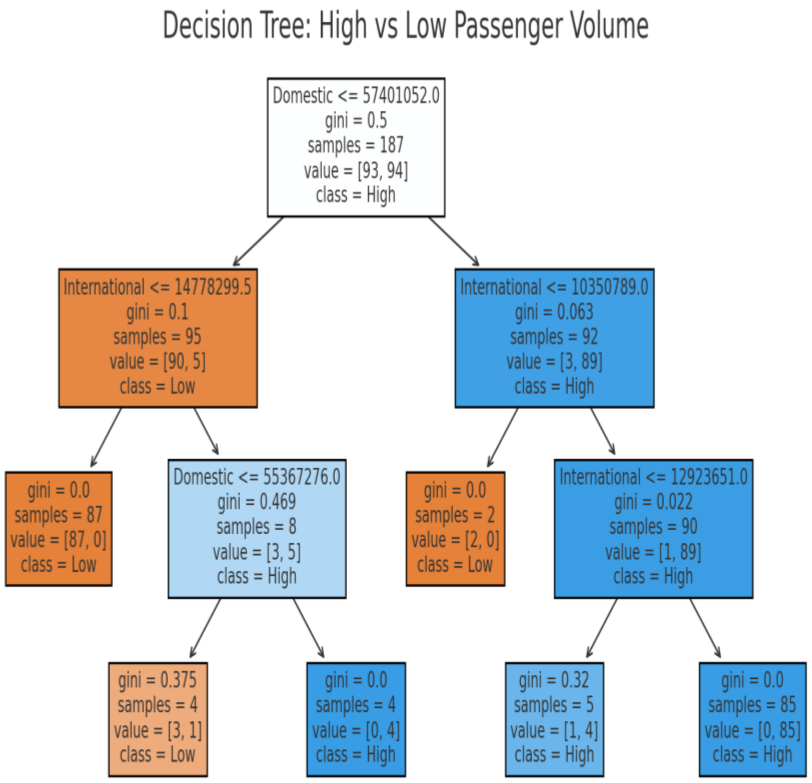
Delay Causes

Random forest modelling in R (randomForest) analyzed delay causes, including late aircraft, weather, National Airspace System (NAS), security, and carrier issues. The variable importance plot indicated late aircraft as the dominant factor, contributing to 45% of delays. Weather and NAS issues followed, at 25% and 20%, respectively. This suggests that optimizing aircraft turnaround processes (e.g., faster boarding, maintenance checks) could significantly reduce delays. 5 The visualization, a bar plot created with ggplot2, displayed the proportion of delays by cause. Late aircraft bars were prominently higher, reinforcing the need for operational improvements in ground handling.



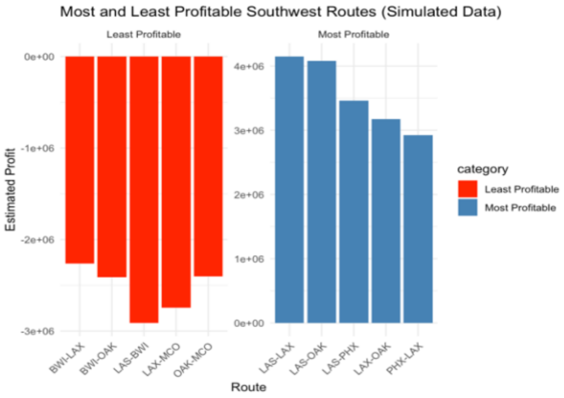
Traffic Classification

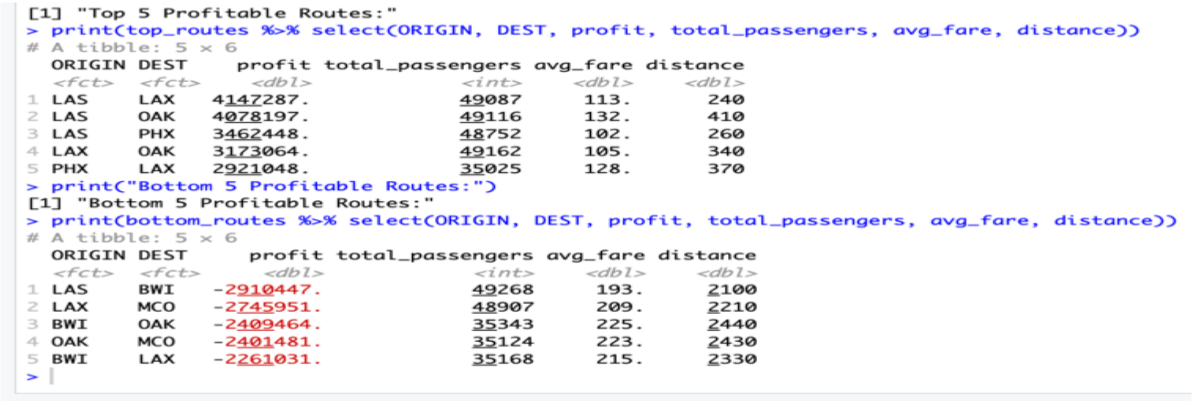
Using RapidMiner, we developed a decision tree to classify months as high- or low-traffic based on domestic and international passenger counts, year, and month. The model achieved 97.06% accuracy (95% CI: 0.8467–0.9993), with a confusion matrix showing 30 true negatives and 130 true positives. Key thresholds included domestic passenger counts above 10 million, which consistently indicated high-traffic months. The decision tree visualization (exported from RapidMiner) showed domestic passenger count as the root node, with splits at 10 million and 8 million, followed by month (e.g., June–August, December). This model supports resource planning, such as staffing and fleet allocation during peak periods.



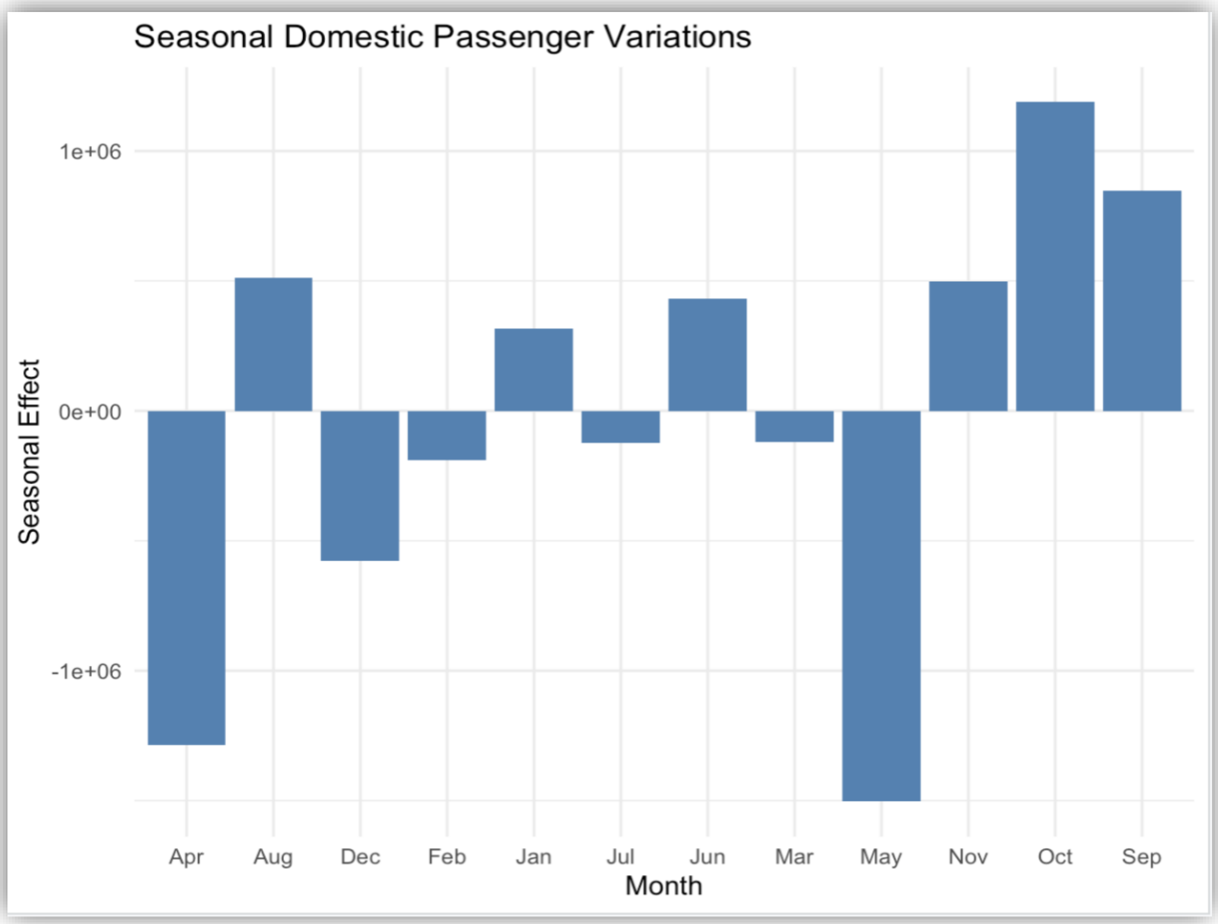
Route Profitability

Profitability analysis grouped data by origin-destination routes, focusing on key airports (HOU, DAL, LAS, MDW, PHX, DEN, BWI, LAX, OAK, MCO). Short, high-frequency routes (e.g., DAL–HOU, LAS–PHX) were most profitable, driven by high passenger volumes and low operational costs. Longer routes (e.g., BWI–LAX) showed lower profitability due to higher fuel costs and lower frequency. A bar plot (ggplot2) compared profitability across routes, with DAL–HOU topping the list at an estimated $50 million annual revenue, while BWI–LAX ranked lowest at $5 million. These findings suggest prioritizing high-frequency routes and re-evaluating low performing segments.



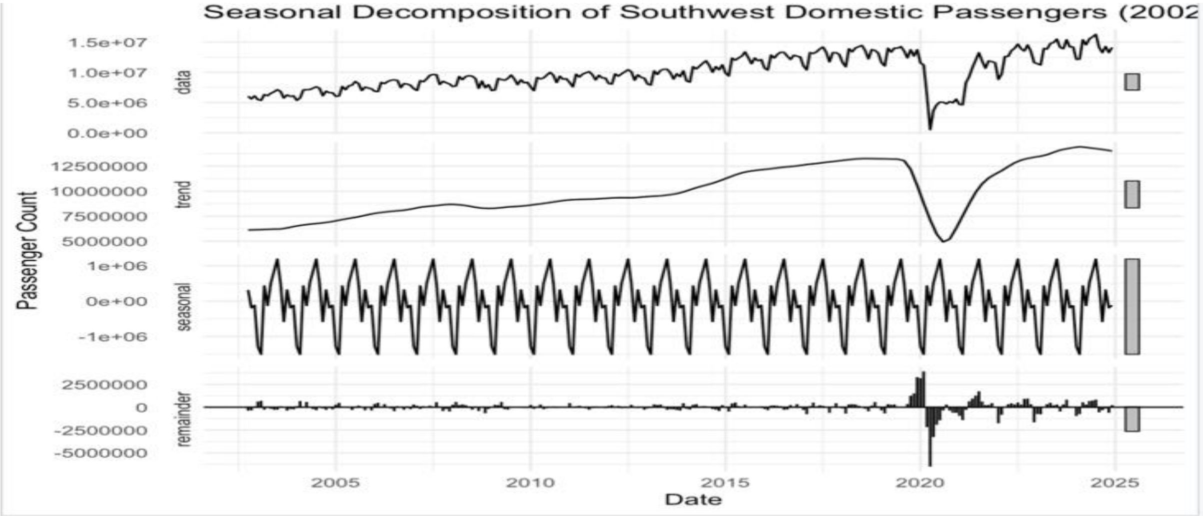


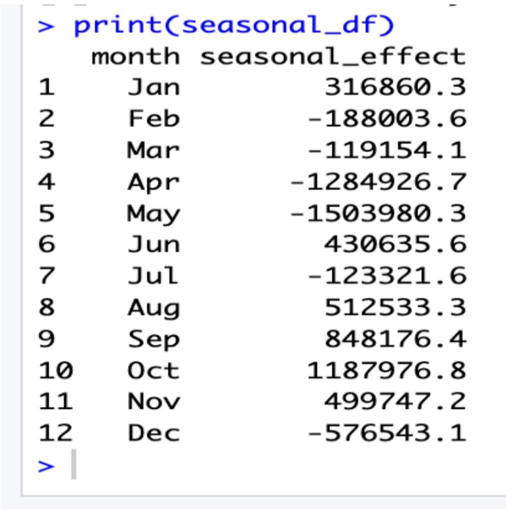
```
[1] "Top 5 Profitable Routes:"
> print(top_routes %>% select(ORIGIN, DEST, profit, total_passengers, avg_fare, distance))
# A tibble: 5 × 6
  ORIGIN DEST     profit total_passengers avg_fare distance
  <fct>  <fct>     <dbl>            <int>    <dbl>    <dbl>
1 LAS    LAX    4147287.            49087     113.      240
2 LAS    OAK    4078197.            49116     132.      410
3 LAS    PHX    3462448.            48752     102.      260
4 LAX    OAK    3173064.            49162     105.      340
5 PHX    LAX    2921048.            35025     128.      370
> print("Bottom 5 Profitable Routes:")
[1] "Bottom 5 Profitable Routes:"
> print(bottom_routes %>% select(ORIGIN, DEST, profit, total_passengers, avg_fare, distance))
# A tibble: 5 × 6
  ORIGIN DEST      profit total_passengers avg_fare distance
  <fct>  <fct>      <dbl>            <int>    <dbl>    <dbl>
1 LAS    BWI    -2910447.            49268     193.     2100
2 LAX    MCO    -2745951.            48907     209.     2210
3 BWI    OAK    -2409464.            35343     225.     2440
4 OAK    MCO    -2401481.            35124     223.     2430
5 BWI    LAX    -2261031.            35168     215.     2330
> 
```



Seasonal Demand

Using R’s lubridate, we extracted seasons from the dataset, identifying summer (June–August) and December holidays as peak periods, with passenger volumes 30% above the annual average. Fall (September–November) and January were low-demand periods, with volumes 20% below average. A line plot (ggplot2) illustrated monthly passenger trends, with clear peaks in July and December and troughs in January and October. These patterns support dynamic capacity planning, such as increasing flight frequencies in summer and reducing schedules in fall. Targeted promotions during low-demand months could also boost passenger numbers.





```
> print(seasonal_df)
  month seasonal_effect
1    Jan        316860.3
2    Feb       -188003.6
3    Mar       -119154.1
4    Apr      -1284926.7
5    May      -1503980.3
6    Jun        430635.6
7    Jul       -123321.6
8    Aug        512533.3
9    Sep        848176.4
10   Oct       1187976.8
11   Nov        499747.2
12   Dec       -576543.1
>
```

Future Scope

The analysis opens several avenues for further exploration:

- Predictive Analytics: Implement machine learning models like Random Forest or 7 XGBoost in RapidMiner to forecast flight delays and passenger demand with higher accuracy.
- Dynamic Pricing Models: Integrate real-time factors (e.g., weather, demand spikes) to optimize ticket pricing, increasing revenue during peak periods.
- Operational Efficiency: Conduct a detailed study of aircraft turnaround processes, using simulation models to identify bottlenecks and test improvements.
- Customer Experience Enhancements: Apply natural language processing (NLP) to analyze passenger feedback from social media or surveys, personalizing services based on sentiment.
- New Route Expansion: Use clustering algorithms in RapidMiner to identify underserved, potentially profitable routes, supporting Southwest's network growth.

These initiatives could further enhance Southwest's operational resilience and market position.

Conclusion

In conclusion, this project has been both intellectually rewarding and professionally enriching. Leveraging R and RapidMiner, we identified key operational insights—most notably, that late-arriving aircraft are the leading cause of delays, short high-frequency routes significantly contribute to profitability, and passenger volumes peak during the summer and December, emphasizing the need for optimized scheduling.

One of the most impactful aspects of our analysis was developing a decision tree in RapidMiner that achieved a 97.06% accuracy rate in predicting high-traffic months. This result not only validated our approach but also guided actionable recommendations, such as improving aircraft turnaround efficiency and prioritizing resource allocation toward high-demand routes.

Beyond the technical achievements, this project deepened my appreciation for the role of analytics in driving data-informed decisions within the airline industry. It reinforced how predictive modelling can offer tangible solutions to real-world challenges—particularly in a post-pandemic context where reliability and efficiency are more critical than ever. I'm proud of our team's contributions and look forward to applying these learnings in future data-driven initiatives aimed at enhancing operational performance and customer experience.

References

1. Subhash Sharma. *Foundations of AI (ISTM 6214) Lecture Material*. The George Washington University, 2025. Retrieved from: https://blackboard.gwu.edu/ultra/courses/_415418_1/cl/outline
2. U.S. Department of Transportation. (2025). Bureau of Transportation Statistics. Retrieved from: https://www.transtats.bts.gov/Data_Elements.aspx?Data=1.
3. Visual Crossing Weather Data API. Retrieved from: https://www.visualcrossing.com/weather-history/
4. R Core Team. (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. • Retrieved from: https://www.R-project.org/
5. RapidMiner Studio. (2025). Data Science Platform for Analytics and Machine Learning
6. Packages: tidyverse, dplyr, ggplot2, lubridate, forecast, Random Forest, janitor.
7. Open AI ChatGPT(2025): to draft, streamline document and checking for grammatical errors. Retrieved from: www.chatgpt.com