Forecasting Air Traffic Demand: Insights and Strategies for the Aviation Industry

Team 1

Andrea Mechery, Kevin Connolly, Chetan Khatri, Shubham Patil, Balaji Prasad Thulasiram

February 19, 2023

OPIM 5671

Table of Contents

[Executive Summary 2](#_90tavz3k3vpl)

[Dataset 3](#_74hc7r92zw3o)

[Final/Best Model 4](#_x79qiumund5v)

[Development 5](#_o4mazmefxf29)

[Exploratory Data Analysis 5](#_blzj2p4ct16)

[Time Series Exploration of RPM 5](#_e55a7pfs3620)

[Modeling Approach 7](#_agg9uwh657u6)

[Forecast Insights 13](#_smv5lz2renn6)

[Business Actions 14](#_8rjeht3g3irr)

[References 15](#_h6bz6zs20m7l)

## Executive Summary

In an era where the aviation industry is marked by ever-changing dynamics, strategic foresight and data-driven decision making are paramount. The process of constructing airplanes is not only time-consuming but also financially demanding, and relies significantly on strategic decisions made by airlines. Navigating the complexities of capacity planning, ticket pricing, and financial sustainability becomes imperative for success, and airlines must grapple with a multitude of considerations. These can include managing aircraft inventory, predicting air travel demand, and optimizing airfare pricing strategies.

This project embarks on a comprehensive exploration of air traffic data for commercial carriers using data provided by the U.S. Department of Transportation, and aims to accurately forecast air travel demand by investigating the Total Revenue Passenger Miles (RPM). RPM was chosen as the primary predictor variable for forecasting air traffic demand due to its direct correlation with passenger activity, which provides stakeholders with a comprehensive measure of both passenger volume and distance traveled, thereby facilitating accurate demand anticipation.

Through a series of modeling approaches, including Seasonal Winters Additive, ARIMA, and ARIMAX models, various factors impacting the target variable were meticulously evaluated. After rigorous analysis and comparison, the final model, the ARIMA(1,1) model on RPM, emerged as the optimal choice for forecasting RPM. This model demonstrated superior accuracy and simplicity compared to alternative approaches.

Using the champion ARIMA (1,1) model, twenty-four months of RPM forecasts were generated to visualize the trajectory. The forecasted values revealed increasing RPMs, thereby indicating an overall increase in air traffic demand. Comparisons were drawn between the annual growth rates observed during pre-pandemic times (Oct 2018 - Sept 2019), the most recent year of available data (Oct 2022 - Sept 2023), and the ensuing forecasted years (Oct 2023 - Sept 2024 & Oct 2024 - Sept 2025). The results unveiled a notably high annual growth rate for the first forecasted year, consistent with the sharp resurgence in RPM immediately following the pandemic. The following forecasted year projects a lower annual growth rate comparable to pre-pandemic levels, indicating the start of stabilization in the aviation sector post-pandemic. These insights from the forecasts can inform several stakeholders in the aviation industry to proactively prepare for the anticipated growth and effectively implement strategic initiatives to accommodate the surge in demand.

## Dataset

The dataset, sourced from the U.S. Department of Transportation’s (DOT) Bureau of Transportation Statistics through Kaggle1, encompasses non-seasonally adjusted monthly air traffic data for all commercial U.S. air carriers, spanning from January 2003 to September 2023. The dataset comprises multiple columns that provide unique perspectives into the aviation industry.

Variables in the dataset include:

1. **Year**: The calendar year in which the data was recorded.
2. **Month**: The specific month within the year for which the data is reported.
3. **Domestic Passengers (Dom\_Pax)**: The number of passengers traveling within the United States.
4. **International Passengers (Int\_Pax)**: The count of passengers traveling internationally.
5. **Total Passengers (Pax)**: The total number of passengers, including both domestic and international travelers.
6. **Domestic Flights (Dom\_Flt)**: The number of flights operating within the United States.
7. **International Flights (Int\_Flt)**: The number of international flights.
8. **Total Flights (Flt)**: The total number of flights, encompassing both domestic and international operations.
9. **Domestic Revenue Passenger-miles (Dom\_RPM)**: The product of the number of domestic passengers and the distance they traveled in thousands.
10. **International Revenue Passenger-miles (Int\_RPM)**: The product of the number of international passengers and the distance they traveled in thousands.
11. **Total Revenue Passenger-miles (RPM)**: The cumulative revenue passenger-miles, combining both domestic and international RPMs.
12. **Domestic Available Seat-miles (Dom\_ASM)**: The product of the number of seats available on domestic flights and the distance flown in thousands.
13. **International Available Seat-miles (Int\_ASM)**: The product of the number of seats available on international flights and the distance flown in thousands.
14. **Total Available Seat-miles (ASM)**: The total available seat-miles of both domestic and international flights.
15. **Domestic Load Factor (Dom\_LF)**: The percentage of revenue passenger-miles as a proportion of available seat-miles for domestic flights (in percent %).
16. **International Load Factor (Int\_LF)**: The percentage of revenue passenger-miles as a proportion of available seat-miles for international flights (in percent %).
17. **Total Load Factor (LF)**: The overall load factor, calculated as the percentage of total revenue passenger-miles relative to total available seat-miles.

The variable selected for forecasting in this project is Total Revenue Passenger-miles (RPM). RPM represents the total number of miles traveled by passengers on commercial flights during a specified period, taking into account both the number of passengers and the distance they travel. Forecasting RPM is of significant importance to the airline industry as revenue serves as a key performance indicator reflecting the demand for air travel, making it a critical metric for assessing the commercial viability and profitability of airline operations.

## Final/Best Model

Four models were analyzed to find an optimal model to forecast Revenue Passenger-miles (RPM). These models are:

1. ARIMA(1,1) Model of RPM.

Table 1 Name = work.AR1\_forecast

1. ARIMAX(2,1) model of RPM incorporating the Ramp variable.

Table 1 Name = work.AR2\_ramp\_forecast

1. ARIMAX(2,1) model of RPM incorporating the Flights (Flt) variable.

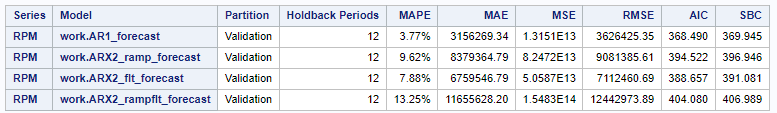
Table 1 Name = work. AR2\_flt\_forecast

1. ARIMAX(2,1) model of RPM incorporating both the Ramp and Flt variables.

Table 1 Name = work.AR2\_rampflt\_forecast

After a thorough comparison of the models on the validation data, the ARIMA(1,1) model emerged as the optimal choice for forecasting Revenue Passenger-miles (RPM). This model exhibited the lowest Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC) values, indicating superior parsimony and simplicity compared to alternative models. Additionally, the ARIMA(1,1) model demonstrated lower Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) values, signifying higher accuracy in predicting RPM values. The selection of this model aligns with the principle of parsimony, favoring simpler models that achieve comparable or better accuracy. Thus, the ARIMA(1,1) model of RPM stands out as the most effective choice to forecast RPM (Table 1).

Table 1: Model Comparison



## Development

### Exploratory Data Analysis

The data cleaning and preprocessing phase began by importing the CSV file obtained from Kaggle to understand its structure. No null values were present in the dataset, however, there were discrepancies in data types due to the presence of commas in numeric values. To address this, all numeric values were converted to the appropriate data type (int64). Additionally, as the dataset contained 'Year' and 'Month' variables, a new variable 'Date' was created by combining 'Year' and 'Month' to adhere to SAS's datetime format requirements. Once these transformations were completed, the dataset was saved to the local drive as a CSV file, ensuring compatibility with SAS for subsequent forecasting analysis.

### Time Series Exploration of RPM

The time series exploration insights for Revenue Passenger-miles (RPM) reveal several significant findings. The analysis, with RPM as the dependent variable and Passenger numbers (Pax), Flights (Flt), Available Seat-miles (ASM), Load Factor (LF), and RAMP as independent variables, was conducted. Seasonality was identified in the data, appearing to follow an additive pattern and a discernible upward trend was observed, suggesting consistent growth over time (Figure 1b & 1d). A major shock to the system occurred in April 2020, coinciding with the onset of the COVID-19 pandemic, leading to a sharp decline in RPM. Despite this, a subsequent recovery was observed, indicating a return to an upward trajectory.

White noise is not present in RPM, so the systematic variation in the data can be used. Furthermore, the autocorrelation function (ACF) analysis demonstrated the presence of autocorrelation in the data, with notable spikes observed at specific lags. Additionally, the partial autocorrelation function (PACF) analysis revealed a spike at lag 1, indicating a significant relationship between RPM and its own lagged values. This suggests that RPM is influenced by its past values, highlighting the importance of incorporating lagged RPM values into forecasting models (Figure 1c).

| **(a)** | **(b)** |
| --- | --- |
| **(c)** | **(d)** |

Figure 1: **(a)** Series Value Plot for RPM **(b)** Seasonal Cycle Plot for RPM **(c)** Correlations Plot for RPM **(d)** Trend Component for RPM

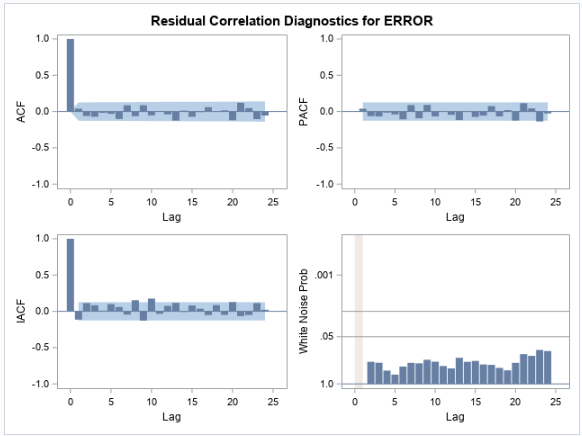
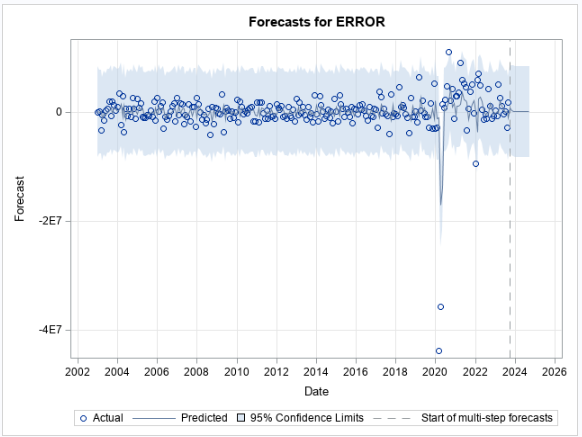
### Modeling Approach

The initial step in the modeling process involved implementing an Additive Seasonal exponential smoothing model since seasonality & trend were present in the time series exploration. The output of this model was assessed through Prediction Error Results, which indicated a failure of the White Noise test (Figure 2b). In other words, the absence of white space indicated that the residuals were exhibiting systematic behavior rather than random variation. This implied that the model was not effectively capturing all relevant information and may not be suitable as a final model for forecasting. This led the group to decide that an ARIMA (AutoRegressive Integrated Moving Average) model would be a more appropriate approach to effectively model the data.

| **(a)** | **(b)** |
| --- | --- |

Figure 2: **(a)** Prediction Errors for RPM **(b)** Prediction Error Correlation for RPM

In the second step of the analysis, an ARIMA (1,1) model was constructed using the errors from the initial Additive Seasonal exponential smoothing model. Examination of the residual correlations through the Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), and Integrated Autocorrelation Function (IACF) revealed no significant spikes, suggesting effective modeling of the data dynamics. Although some white noise was detected, the incorporation of error forecasting in the ARIMA model would allow for improved accuracy by capturing any remaining unexplained variation in the data. However, this approach was ultimately not considered in the final model evaluations. It was recognized that incorporating the error forecasting into the ARIMA model could introduce unnecessary complexity beyond the scope of this project. As a result, the decision was made to focus on more straightforward modeling techniques to ensure feasibility and practicality within the project constraints.



**(a)** **(b)**

Figure 3: **(a)** Forecast for ESM Errors **(b)** Prediction Error Correlation for RPM

In the third step of the modeling process, an ARIMA (AutoRegressive Integrated Moving Average) model was applied to the RPM variable. The chosen ARIMA configuration consisted of parameters p=1, d=1, and q=0. Additionally, a seasonal ARIMA model with parameters P=0, D=1, and Q=1 was used to account for seasonal patterns in the data. The output of the ARIMA modeling revealed some white noise present in the residuals (Figure 4b), indicating that the model effectively captured the systematic variation in the data. Furthermore, no significant spikes were observed in the autocorrelation function (ACF), partial autocorrelation function (PACF), or integrated autocorrelation function (IACF), suggesting that the model adequately accounted for autocorrelation and was suitable for forecasting the RPM variable.

| **(a)** | **(b)** |
| --- | --- |

Figure 4: **(a)** Forecast for RPM using ARIMA (1,1) model **(b)** Residual Correlation Diagnostics Plot of RPM using ARIMA (1,1) model

In the fourth step of the exploration process, a Prewhitening Model was implemented to further investigate the relationship between the RPM variable and its potential influencing factors. The objective was to determine the accuracy and lag at which each independent variable (such as Pax, ASM, LF, and Flt.) impacts the RPM, thereby providing information whether to include any relevant variables in subsequent modeling iterations. The output of the Cross Correlation analysis (Figure 5b) revealed that Flights (Flt) exhibited a lag 1 effect on RPM, suggesting a significant influence on RPM at a one-month lag. In contrast, Pax, ASM, and LF did not demonstrate any significant effects on RPM at the examined lags (Figures 5a, 5c, and 5d). As a result, the group decided to incorporate Flights as a predictor variable in the next modeling phase to assess its impact on driving and improving forecasting accuracy.

| **(a)** | **(b)** |
| --- | --- |
| **(c)** | **(d)** |

Figure 5: **(a)** Cross Correlations of RPM and ASM **(b)** Cross Correlations of RPM and Flt **(c)** Cross Correlations of RPM and LF **(d)** Cross Correlations of RPM and Pax

In the fifth step of the modeling process, an ARIMAX (AutoRegressive Integrated Moving Average with Exogenous variables) model was constructed, incorporating an intervention dummy variable called Ramp to account for the impact of the COVID-19 pandemic. The Ramp variable was created by assigning zeroes to normal periods and linearly spaced negative values during the pandemic-induced downturn. The ARIMA configuration consisted of parameters p=2, d=1, and q=0, and a seasonal ARIMA model with parameters P=0, D=1, and Q=1 was included, with Ramp serving as the exogenous variable. The White Noise Probability plot indicates that the signal in the original series was nearly completely explained by the model’s parameters. What remains (the residuals) seem to be white noise. However, no significant spikes were observed in the autocorrelation function (ACF), partial autocorrelation function (PACF), or integrated autocorrelation function (IACF), suggesting that the model adequately addressed autocorrelation (Figure 6).

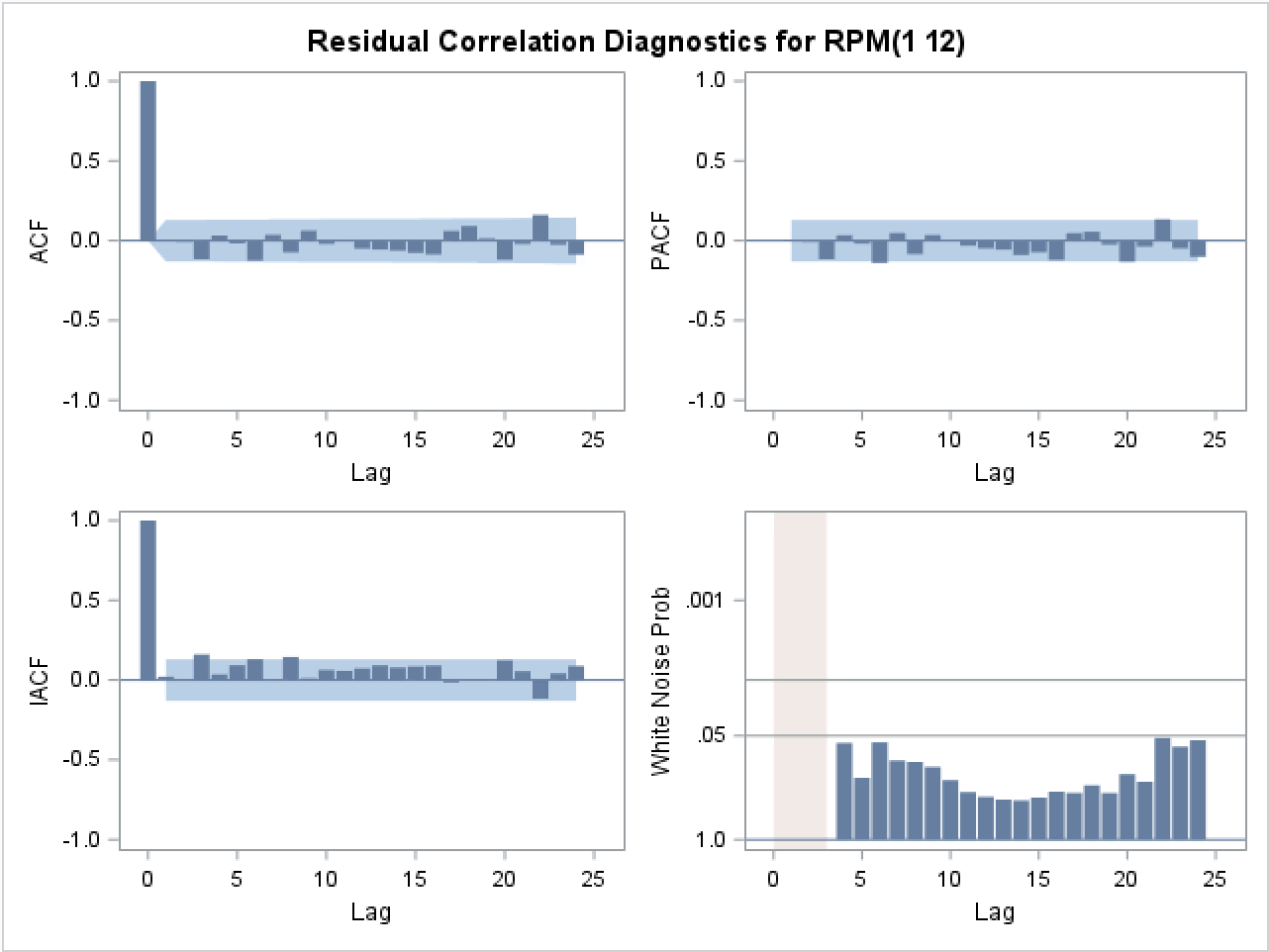
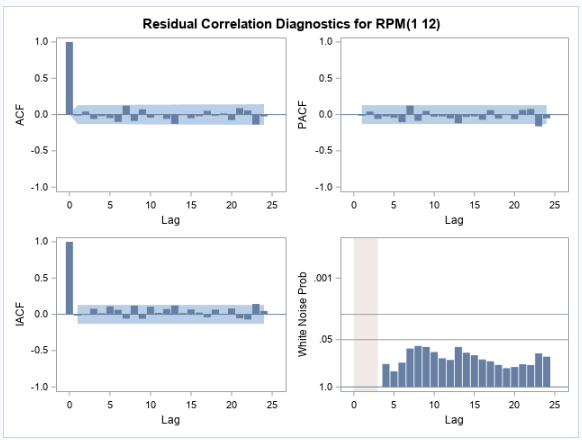
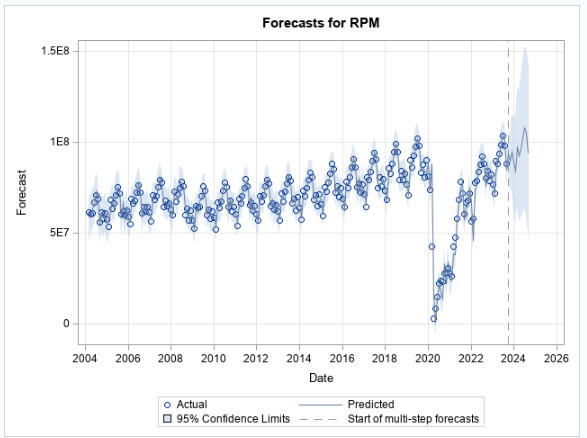


Figure 6: Residual Correlation Diagnostics for RPM using ARIMAX (2,1) model with Ramp as the X Variable

In the sixth step of the analysis, an ARIMA model was applied to the Flights (Flt) variable to explore its potential impact on forecasting RPM. Flt was selected for investigation based on its observed lag 1 effect on RPM in the prewhitening model's cross-correlation analysis. The objective of this step was twofold: firstly, to forecast 12 months of Flt data independently; and secondly, to incorporate Flt as an independent variable in an ARIMAX model alongside Ramp, thus providing a comprehensive dataset for forecasting RPM values for the following 12 months. By understanding how Flt interacts with RPM and incorporating it into the forecasting models, the aim was to enhance the accuracy and robustness of the predictive model for future time periods.

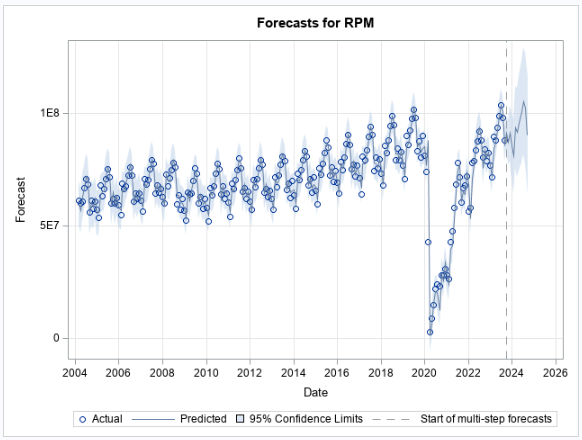
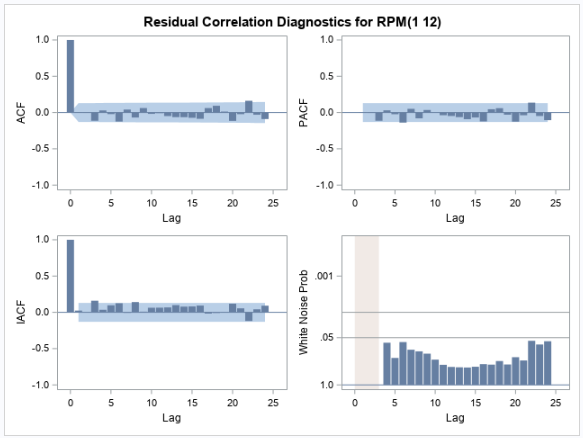
In the seventh step of the analysis, an ARIMAX model was developed using Flights (Flt) as the exogenous variable to enhance the forecasting accuracy of RPM. The ARIMA configuration consisted of parameters p=2, d=1, and q=0, and a seasonal ARIMA model with parameters P=0, D=1, and Q=1 was included. Notably, Flt was included as the independent variable with a lag 1 effect, derived from the observed lag 1 effect on Flt identified in the prewhitening analysis. By incorporating this lag effect, the model's residual correlations significantly improved, resulting in a more reliable forecast. Furthermore, the white noise probability chart indicated favorable results when the lag 1 effect was incorporated and no significant autocorrelation was detected in the residuals, further validating the model (Figure 7).



**(a) (b)**

Figure 7: **(a)** Forecast for RPM using ARIMAX (2,1) model **(b)** Residual Correlation Diagnostics Plot of RPM using ARIMAX (2,1) model

In the eighth step of the analysis, an ARIMAX model was developed using Flights (Flt) and Ramp (Covid) as the exogenous variables to enhance the forecasting accuracy of RPM. The ARIMA configuration consisted of parameters p=2, d=1, and q=0, and a seasonal ARIMA model with parameters P=0, D=1, and Q=1 was included. Flt was also included as an independent variable with a lag 1 effect. Furthermore, the white noise probability chart indicated favorable results and no significant autocorrelation was detected in the residuals, further validating the model (Figure 8).

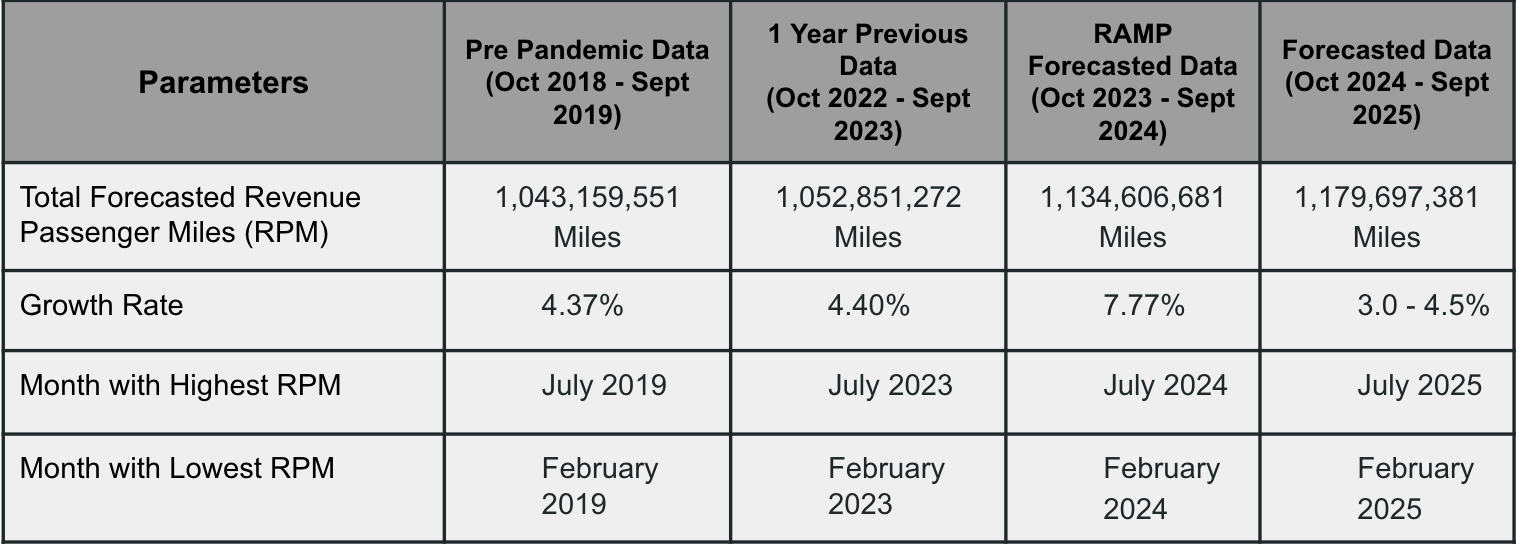
**(a) (b)**

Figure 8: **(a)** Forecast for RPM using ARIMAX (2,1) model **(b)** Residual Correlation Diagnostics Plot of RPM using ARIMAX (2,1) model

In the final step of the analysis, an Out-of-Sample model was created to evaluate the forecasting performance of the previously created models. This procedure involved partitioning the data into two parts: a fitting dataset and a holdout dataset, with a holdback period of 12 months. Four models were trained and evaluated on both the fitting and holdout datasets. These models included an ARIMA(1,1) model on RPM, an ARIMAX(2,1) model on RPM incorporating the Ramp variable, an ARIMAX(2,1) model on RPM incorporating the Flights (Flt) variable, and an ARIMAX(2,1) model on RPM utilizing both the Ramp and Flt variables as exogenous predictors. The accuracy of each model was then assessed to determine its effectiveness in forecasting RPM values.

## Forecast Insights and Business Actions

The best model is the ARIMA(1,1) model because, as we know, RPM is an independent variable with insignificant impact from other variables in the dataset. Despite being the simplest model, it performs the best. The growth impacts derived from this forecast model is summarized in Table 2.

Table 2: Yearly comparison of Actual & Forecasted RPMs

Pre-pandemic growth rate for the industry was found to be roughly 4.4% year over year. It was observed that the pandemic rebound growth rate during the ramp was to the order of 7.7% year over year. The forecasted growth and what should be used to guide business actions post-pandemic was forecasted to be between 3 and 4.5%; which is similar to pre-pandemic growth. Because of the seasonality of the dataset, we can see that, throughout the four time periods, February has the lowest RPM and July has the greatest RPM.

As the aviation industry rebounds from the pandemic, forecasts indicate a return to pre-pandemic growth rates. For airports, this suggests maintaining momentum on expansion and renovation plans postponed during the pandemic, and prioritizing workforce growth, particularly in critical bottleneck positions. Airlines should proceed with fleet expansion and upgrades to meet increasing demand, ensuring budget availability and expediting airplane orders to align with the recovering supply chain. Additionally, adjusting ticket pricing strategies to accommodate forecasted growth is essential for maximizing revenue in the evolving market landscape.

## 

## References

1 <https://www.kaggle.com/datasets/yyxian/u-s-airline-traffic-data>