ETW3420 Group Assignment

Tutorial 1 (Friday, 2pm)

Group members:

Daphney Loka (32685327)
Madeline Ng Yan Mei (32935412)
Koay Zi Yan (30882303)

Table of Content

Introduction

The latest pandemic is a human tragedy, which has affected the lives of hundreds and thousands of people while leaving a growing impact on the global economy (Naseer et al., 2023). With governments around the world enforcing a long period of lockdown in their country for the safety of the public, the aviation industry has been the one most affected (Salas, 2022). Airlines, airports, and the broader air travel ecosystem have faced numerous challenges and disruptions, necessitating a careful and data-driven analysis of the consequences. The US airlines have experienced substantial losses during the pandemic due to flight restrictions and border lockdowns resulting in an abrupt decline in passenger traffic (Warnock-Smith et.al, 2021).

In this report, we will be undertaking the task of assessing the impact of the pandemic on the US Revenue Passenger Miles (RPM). The RPM metric serves as a key indicator of passenger demand and economic health within the aviation industry. This report aims to utilize the Box-Jenkins methodology, specifically the ARIMA framework, to provide a quantitative analysis of the loss in US RPM since the pandemic.

The objectives of this analysis are: first, to understand the trajectory of RPM before, during, and after the pandemic; second, to quantify the extent of loss experienced by the US aviation industry; third, to forecast potential future trends and recovery paths for this critical economic indicator. We will identify the effect of the pandemic on the aviation industry in the United States by assessing the air traffic statistics for the country using forecasting methods.

This report is a comprehensive and data-driven investigation, employing the capabilities of the R studio to facilitate a thorough understanding of the aviation sector's response to the Covid-19 pandemic. It seeks to provide stakeholders, policymakers, and the public with valuable insights and understanding of the impact on US RPM, thereby contributing to informed decision-making and the development of strategies to address the challenges that persist in the post-pandemic landscape.

Phase 1: Model Identification

Data Preparation

Based on our graphical visualisation analysis, we chose not to use the whole dataset. This is due to the breaks in the time plot caused by the incident of 9/11 and The Great Recession. Hence, considering the dataset from 2010 should give us better forecasts which are a clear representation of the past values as there were no breaks since that period. We decided to remove the previous years and consider only the period from 2010-January until 2023-April to use for our prediction modelling.

With our data finalised, we divide it into two potions. The first portion ("Pre-Covid" data) is comprised of the period before the pandemic hit the US, starting from January 2010 until January 2020, while the second portion ("Covid" data) is comprised of the period after the pandemic hit the US, starting from February 2020 and end on July 2021. We will be using autoregressive integrated moving average, or ARIMA model to predict and create forecasts for Covid period (Phase 3) based on the Pre-Covid dataset (Phase 2).

Splitting helps to avoid overfitting and to improve the training dataset accuracy. Hence, we will partition the dataset into a training set and a test set using 70-30 ratio. Below we have attached the plot of training and test partition of pre-covid data shown in **Figure 1.1**. Our training dataset consists of data points starting from January 2010 until January 2017. Whereas our test set consists of data points starting from February 2017 to January 2020. Hence, the training set has 85 observations and the test set has 36 observations.

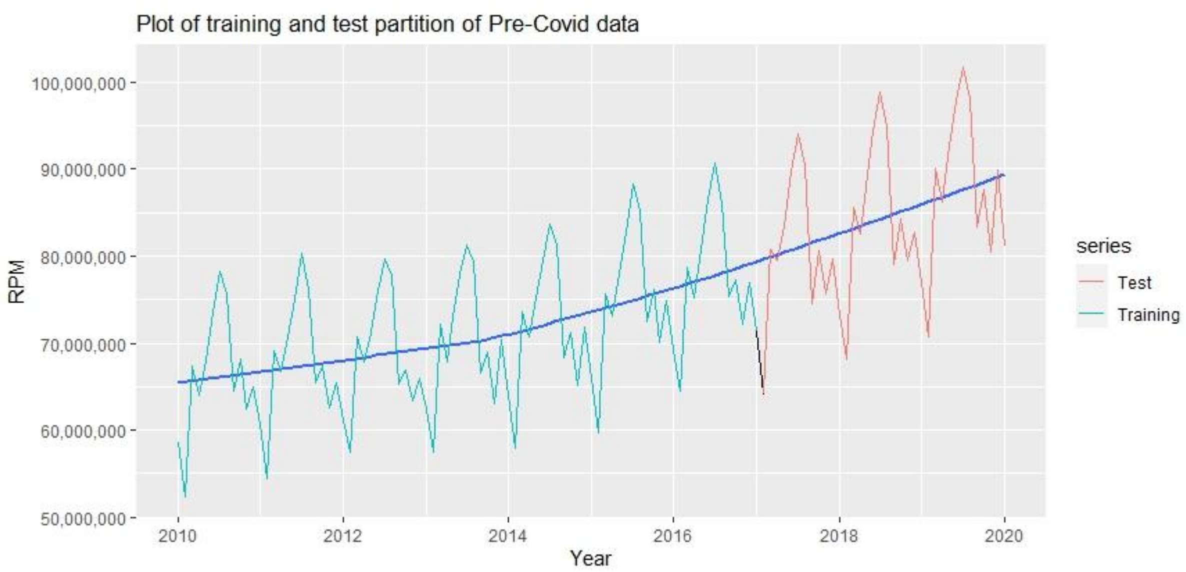

Figure 1.1: Training and Test Partition of Pre-Covid data Plot

We will employ the training set to build ARIMA models and utilize the test set to evaluate and compare their accuracy, and is plotted as below.

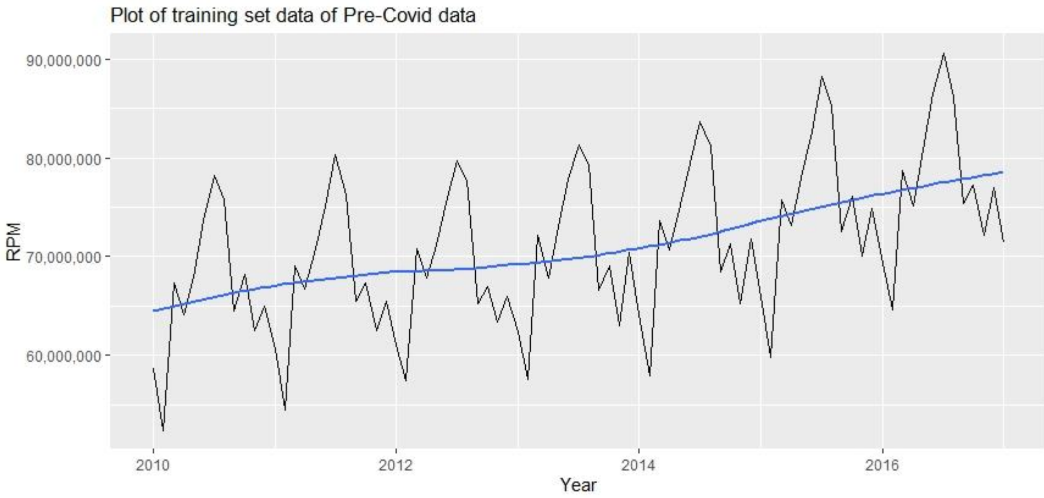

Figure 1.2: The Training set of Pre-Covid ata Plot

Before identifying possible ARIMA models and applying them, we have to ensure that our time series is stationary by stabilising the variance and mean using transformation and differencing. The initial visual inspection of our time series data showed a quite constant variance, an upward trend, and a clear seasonality pattern as it is a monthly dataset. But to ensure whether a transformation is needed to stabilise the variance, we will calculate the Lamba. Our Lambda result is 0.462, meaning that the variance increases slightly over time, hence our first step is to apply Box-Cox transformation to stabilise the variance. The seasonality and trend indicate that the mean is not constant (see Figure 1.2), hence differencing is needed. We will first apply seasonal differencing to remove the strong seasonality effect and as an attempt to stabilise the mean.

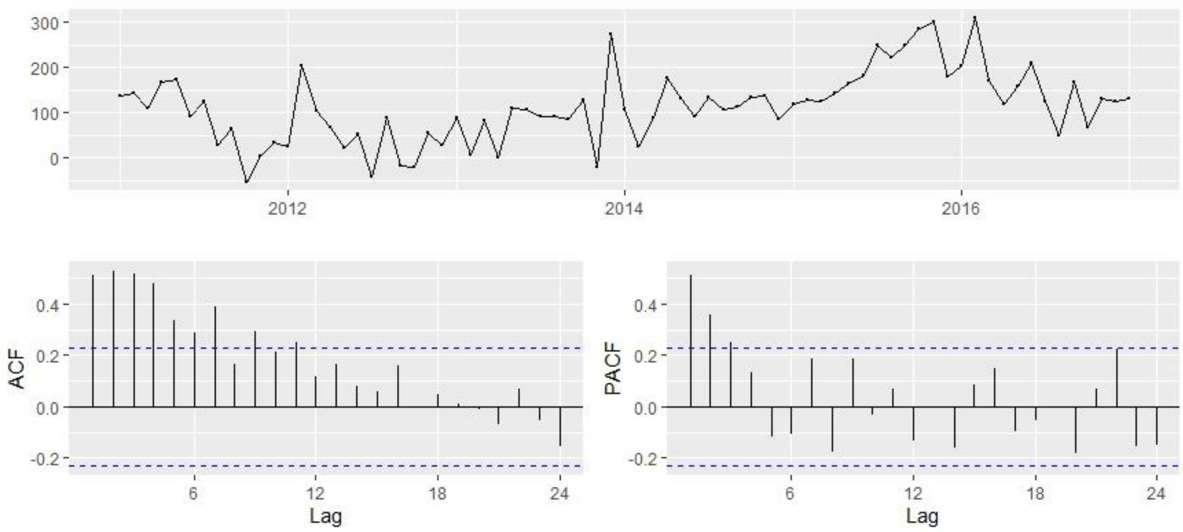

Figure 1.3: Plot after performing seasonal differencing

The seasonality differenced data is shown in **Figure 1.3**, but it still appears to be non-stationary as the mean is still not constant and there are many spikes in ACF above the blue line (significance threshold). Moreover, we also perform ndiffs() function to test whether we need another first differencing and the result of [1] shows that it is required.

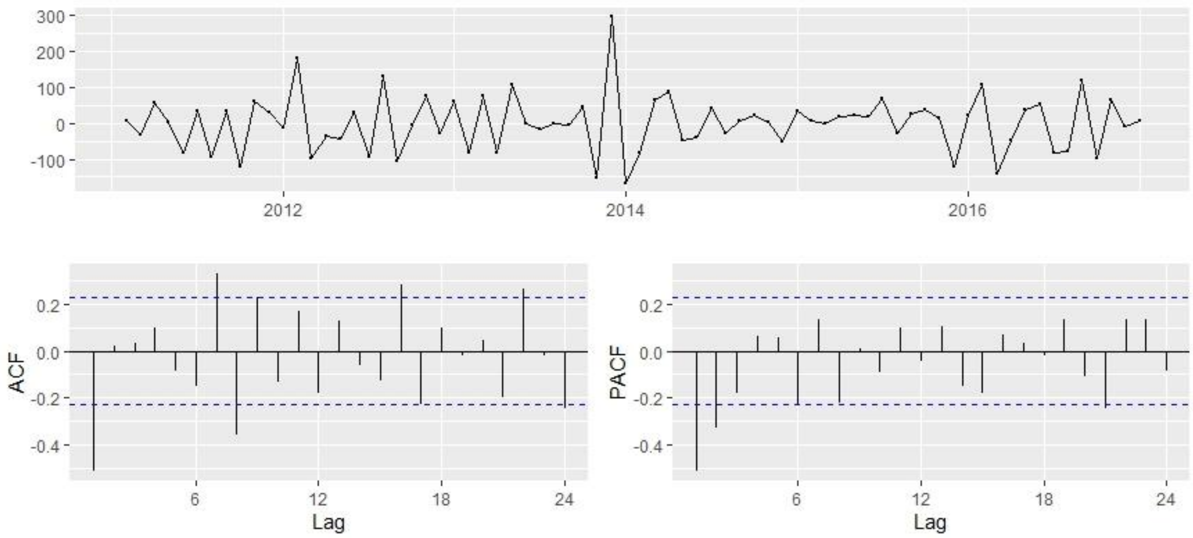

Figure 1.4: Plot after performing seasonal differencing and first differencing

After performing the first difference (see Figure 1.4), the mean seems to be constant and the result ndiffs() function also shows no additional differencing is needed anymore.

<u>Identify ARIMA Models</u>

We will manually identify our ARIMA models using this format ARIMA $(p, d, q)(P, D, Q)_m$, where the lowercase notations represent the non-seasonal part of the model, while uppercase notations represent the seasonal part. Since we require first differencing and seasonality differencing to achieve a stationary time series, d and D will both be 1 respectively. The notation m here indicates the number of observations per year and since our time series is a monthly data, our m will be 12.

Hence, our ARIMA format will be ARIMA $(p, 1, q)(P, 1, Q)_{12}$. Using this format, we have identified four model variations based on the ACF and PACF plot in Figure 1.4. Additionally, we include constant (c = 0) in all our models because we want the slope to follow the past observations which contain a trending behaviour rather than a straight line slope equal to the mean of the differenced data.

#1: p and P = 0

The first ARIMA model will consider that p and P are 0 (ARIMA $(0, 1, q)(0, 1, Q)_{12}$) hence we will be only identifying the q and Q by looking at the ACF plot. We observed a significant spike in the first lag in ACF, suggesting a non-seasonal MA(1) component where q =1. Given the seasonality of the data set, there is a small spike in lag 24 and the line in lag 24 is almost over the blue line, hence we consider that Q = 2 to account for seasonal patterns or dependencies in the data making a seasonal MA(2) component. Hence our Model 1 is identified as ARIMA $(0, 1, 1)(0, 1, 2)_{12}$.

#2: Variation from Model 1

The next model is the variation from Model 1 considering that there are two significant spikes at lags 1 and 2 in PACF. Hence p is no longer 0 but is 2 hence suggesting a non-seasonal AR(2) model. Therefore, our second ARIMA model is identified as ARIMA $(2, 1, 1)(0, 1, 2)_{12}$.

#3: q and Q = 0

Our third model will assume that q and Q are 0 (ARIMA $(p, 1, 0)(P, 1, 0)_{12}$ hence we will be only identifying the p and P by looking at the PACF plot. Two significant spikes at non-seasonal lags (lag 1 and 2) are noticed in PACF which suggests a non-seasonal AR(2) component where p = 2. Furthermore, there are no significant spikes observed in the seasonality lags (lag 12 or 24) in the PACF plot, hence there is no moving average component and P = 0. Therefore our third model is identified as ARIMA $(2, 1, 0)(0, 1, 0)_{12}$.

#Variation from Model 3

Our last model is a variation from Model 3 whereby we include Q = 2 as we consider the existence of seasonal lag spikes in ACF to address seasonal dependencies in the data whereby there is a small spike in lag 24 and the line in lag 24 is almost over the blue line. Hence, by considering a seasonal MA(2) component our final model becomes ARIMA $(2, 1, 0)(0, 1, 2)_{12}$.

Hence, our final four models are ARIMA $(0, 1, 1)(0, 1, 2)_{12}$, ARIMA $(2, 1, 1)(0, 1, 2)_{12}$, ARIMA $(2, 1, 0)(0, 1, 0)_{12}$, and ARIMA $(2, 1, 0)(0, 1, 2)_{12}$

<u>Phase 2: Estimation and Testing</u>

In addition to the four manually identified ARIMA models in Phase 1, we will identify one more ARIMA model by using auto.arima() function in R as our fifth model. An additional argument of (stepwise = FALSE, approximation = FALSE) is also added which lets R run a more intensive search by considering every possible combination of ARIMA orders and ensuring that R can accurately compute the best model, although it takes more time to load. As a result, the best model computed by R is ARIMA$(0, 1, 1)(2, 1, 0)_{12}$, hence it becomes our fifth model as it does not collide with any of our previous models. Afterwards, we will generate the parameter estimates of all of the ARIMA models that we have identified shown in **Table 2.1**.

Table 2.1: Parameter Estimates of all the identified ARIMA models

ARIMA(p, q, d)(P, Q, D)	AR(p)		MA(q)		Seasonal AR(P)		Seasonal MA(Q)	
	AR(1)	AR(2)	MA(1)	MA(2)	Seasonal AR(1)	Seasonal AR(2)	Seasonal MA(1)	Seasonal MA(2)
ARIMA $(0, 1, 1)(0, 1, 2)_{12}$	-	-	-0.631	-	-	-	-0.578	-0.422
ARIMA $(2, 1, 1)(0, 1, 2)_{12}$	-0.228	-0.040	-0.458	-	-	-	-0.547	-0.453
ARIMA $(2, 1, 0)(0, 1, 0)_{12}$	-0.675	-0.322	-	-	-	-	-	-
ARIMA $(2, 1, 0)(0, 1, 2)_{12}$	-0.650	-0.255	-	-	-	-	-0.537	-0.463
ARIMA $(0, 1, 1)(2, 1, 0)_{12}$	-	-	-0.611	-	-0.232	-0.372	-	-

By inserting the parameters shown in Table 2.1, we can write the ARIMA equation for each model as below:

Model 1 – ARIMA$(0, 1, 1)(0, 1, 2)_{12}$

$(1 - B)(1 - B^{12})y_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})(1 + \Theta_2 B^{24})\, \varepsilon_t$

$(1 - B)(1 - B^{12})y_t = (1 - 0.631\, B)(1 - 0.578\, B^{12})(1 - 0.422\, B^{24})\, \varepsilon_t$

Model 2 – ARIMA$(2, 1, 1)(0, 1, 2)_{12}$

$(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - B^{12})y_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})(1 + \Theta_2 B^{24})\, \varepsilon_t$

$(1 - (-\, 0.228)B - (-\, 0.040)B^2)(1 - B)(1 - B^{12})y_t = (1 - 0.458\, B)(1 - 0.547\, B^{12})(1 - 0.453\, B^{24})\, \varepsilon_t$

$(1 + 0.228\, B + 0.040\, B^2)(1 - B)(1 - B^{12})y_t = (1 - 0.458\, B)(1 - 0.547\, B^{12})(1 - 0.453\, B^{24})\, \varepsilon_t$

Model 3 – ARIMA$(2, 1, 0)(0, 1, 0)_{12}$

$(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - B^{12})y_t = \varepsilon_t$

$(1 - (-\, 0.675)B - (-\, 0.322)B^2)(1 - B)(1 - B^{12})y_t = \varepsilon_t$

$(1 + 0.675\, B + 0.322\, B^2)(1 - B)(1 - B^{12})y_t = \varepsilon_t$

Model 4 – ARIMA$(2, 1, 0)(0, 1, 2)_{12}$

$(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - B^{12})y_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})(1 + \Theta_2 B^{24})\,\varepsilon_t$

$(1 - (-0.650)B - (-0.255)B^2)(1 - B)(1 - B^{12})y_t = (1 - 0.537\,B^{12})(1 - 0.463\,B^{24})\,\varepsilon_t$

$(1 + 0.650\,B + 0.255\,B^2)(1 - B)(1 - B^{12})y_t = (1 - 0.537\,B^{12})(1 - 0.463\,B^{24})\,\varepsilon_t$

Model 5 (auto.arima()) – ARIMA$(0, 1, 1)(2, 1, 0)_{12}$

$(1 + \Phi_1 B^{12})(1 + \Phi_2 B^{24})(1 - B)(1 - B^{12})y_t = (1 + \theta_1 B)\,\varepsilon_t$

$(1 - (-0.232)B^{12})(1 - (-0.372)B^{24})(1 - B)(1 - B^{12})y_t = (1 - 0.611\,B)\,\varepsilon_t$

$(1 + 0.232\,B^{12})(1 + 0.372\,B^{24})(1 - B)(1 - B^{12})y_t = (1 - 0.611\,B)\,\varepsilon_t$

Rank the ARIMA models

AIC, BIC, and AICc are the three common metrics used for model comparison and selection. In our case, we will be using AICc as it is more suitable when the sample size is small. The ARIMA models will be ranked based on the lowest AICc and we will select the top 2 best models. As shown in **Table 2.2**, we can see that Model 1 and Model 4 have the two lowest AICc values, hence we will use these two models in the next step.

Table 2.2: AICc value of all five ARIMA models

	Model	AICc	Rank
Model 1	**ARIMA**$(0, 1, 1)(0, 1, 2)_{12}$	**796.233**	**#1**
Model 2	ARIMA$(2, 1, 1)(0, 1, 2)_{12}$	800.234	#3
Model 3	ARIMA$(2, 1, 0)(0, 1, 0)_{12}$	808.504	#5
Model 4	**ARIMA**$(2, 1, 0)(0, 1, 2)_{12}$	**799.037**	**#2**
Model 5	ARIMA$(0, 1, 1)(2, 1, 0)_{12}$	801.737	#4

Residual Diagnostic for the top 2 models

After selecting the two best models based on AICc, it is important to check if the residuals in each of the model is white noise using the checkresiduals() function. Having a white noise means that the model is able to capture the data changes effectively and hence determines a reliable forecasting model.

1. **ARIMA**$(0, 1, 1)(0, 1, 2)_{12}$

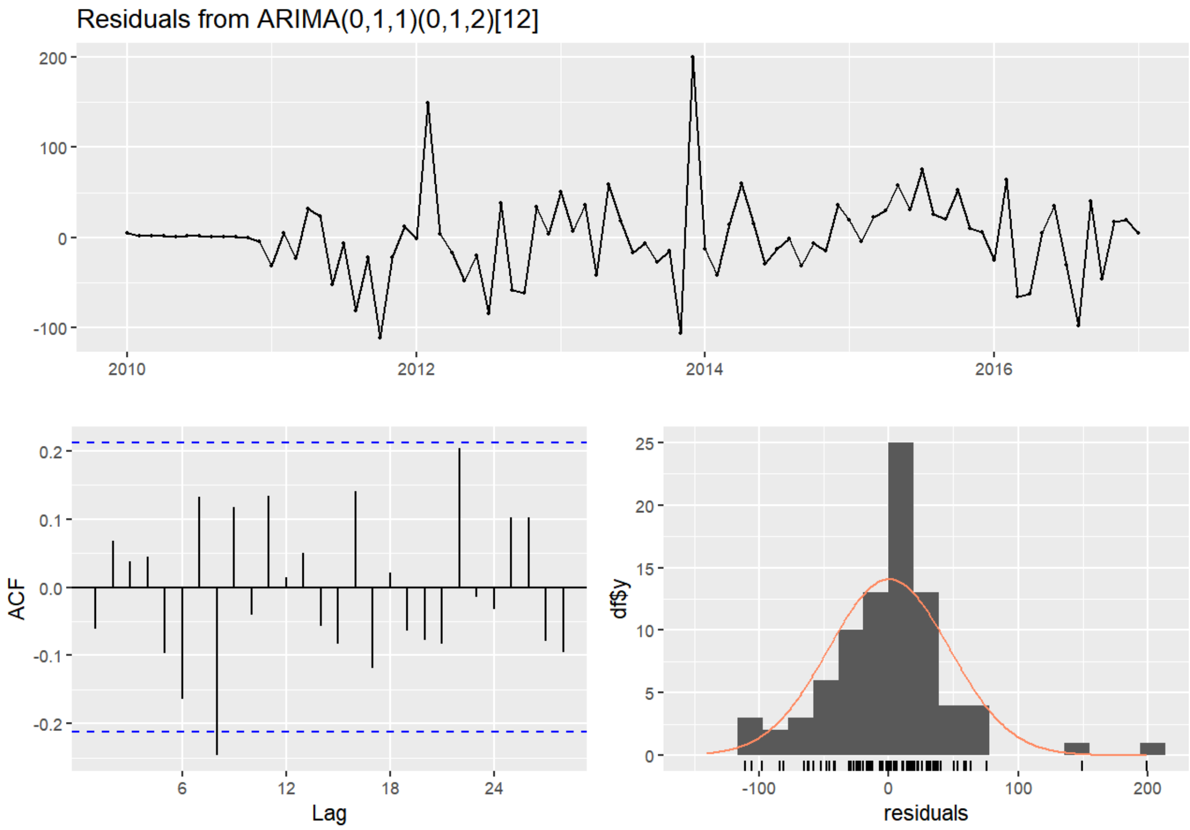

Figure 2.1: Residuals from **ARIMA**$(0, 1, 1)(0, 1, 2)_{12}$ model

<u>Step 1:</u>

$H_0 = \rho_1 = \rho_2 = \ldots\ldots = \rho_{17} = 0$

H_1: at least one of $\rho_i \neq 0$ for $i = 1, 2, \ldots\ldots, 17$

<u>Step 2:</u>

Significance level (α) = 0.05

<u>Step 3:</u>

p-value = 0.122

<u>Step 4:</u>

Decision Rule:

The p-value (0.122) is bigger than the 5% significance level, hence, we do not reject the null hypothesis (H_0). We can say that the residual in this model is white noise on 5% significance level since . In conclusion, ARIMA$(0, 1, 1)(0, 1, 2)_{12}$ model is adequate and reliable to be used in forecasting.

2. **ARIMA**$(2,\ 1,\ 0)(0, 1, 2)_{12}$

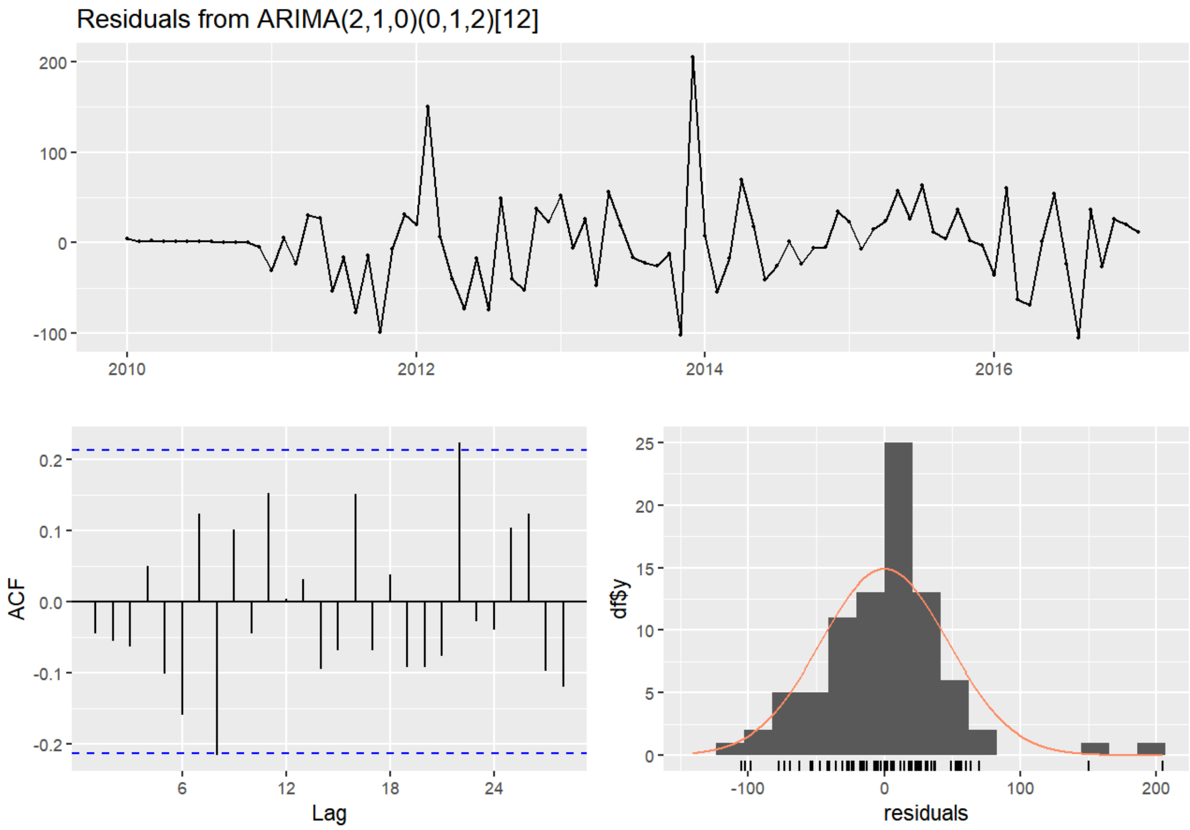

Figure 2.2: Residuals from **ARIMA**$(0, 1, 1)(0, 1, 2)_{12}$ model

<u>Step 1:</u>

$H_0 = \rho_1 = \rho_2 = \ .....\ = \rho_{17} = 0$

H_1: at least one of $\rho_i \neq 0$ for $i = 1,\ 2,\ .....\ ,\ 17$

<u>Step 2:</u>

Significance level (α) = 0.05

<u>Step 3:</u>

p-value = 0.149

<u>Step 4:</u>

Decision Rule:

The p-value (0.149) is bigger than the 5% significance level, hence, we do not reject the null hypothesis (H_0). We can say that the residual in this model is white noise on 5% significance level since . In conclusion, ARIMA$(2,\ 1,\ 0)(0, 1, 2)_{12}$ model is adequate and reliable to be used in forecasting.

Does re-identification required?
The two top models passed the required checks of being reliable forecasting models as both residuals are white noise in 5% significance level. Additionally, it also fulfils the criteria of being white noise in 10% significance level as the p-value for both models is greater than 0.1. Therefore, we are satisfied with the model selection and no re-identification is required anymore.

The best forecasting model
To identify the best forecasting model, we will compare the top 2 models to see which model has the lowest AICc. Model 1 is the best forecasting model and will be used in Phase 3.

Table 2.2: AICc Comparison between the Top 2 Models

Model		**AICc**
Model 1	ARIMA$(0, 1, 1)(0, 1, 2)_{12}$	796.233
Model 4	ARIMA$(2, 1, 0)(0, 1, 2)_{12}$	799.037

The backshift notation of the best model, which is Model 1 – ARIMA$(0, 1, 1)(0, 1, 2)_{12}$, can be written as such:

$$(1 - B)(1 - B^{12})y_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})(1 + \Theta_2 B^{24})\,\varepsilon_t$$

$$(1 - B)(1 - B^{12})y_t = (1 + 0.631\,B)(1 + 0.578\,B^{12})(1 + 0.422\,B^{24})\,\varepsilon_t$$

Phase 3: Application
From Phase 2, we have selected the best ARIMA model which is ARIMA$(0,1,1)(0,1,2)_{12}$. Using the selected ARIMA model, we will re-estimate it with Pre-Covid data to obtain the estimation model and produce forecasts for the Covid time period which is shown in **Figure 3.1** and parameter estimates in **Table 3.1** below.

Table 3.1: Parameter estimates of ARIMA$(0,1,1)(0,1,2)_{12}$ model on "Pre-Covid" data

	MA(1)	Seasonal MA(1)	Seasonal MA(2)
ARIMA$(0,1,1)(0,1,2)_{12}$	-0.7020	-0.4779	-0.1913

The backshift notation of the estimated model can be written as such:

$$(1 - B)(1 - B^{12})y_t = (1 - 0.702B)(1 - 0.4779B^{12})(1 - 0.1913B^{24})\varepsilon_t$$

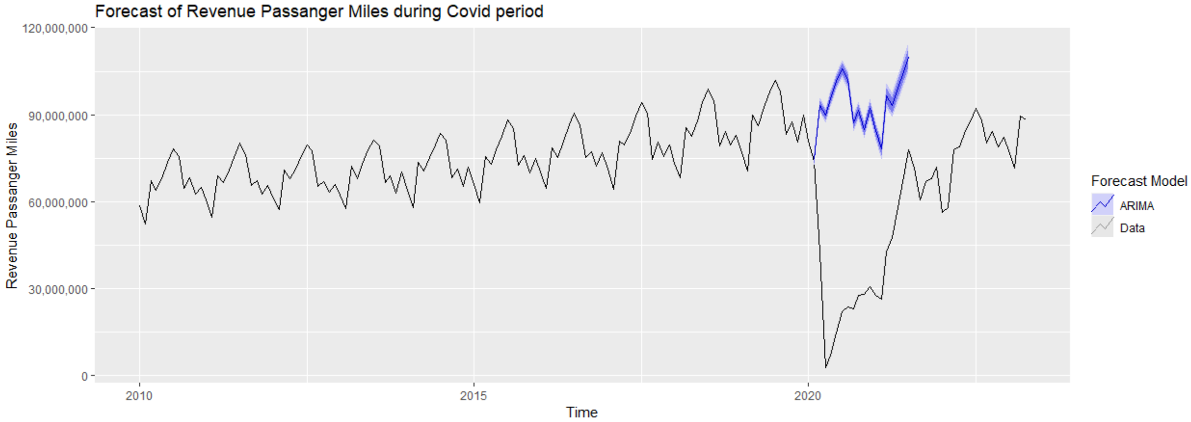

Figure 3.1: Forecast for "Covid" time period using ARIMA model

The line graph shows the forecast of revenue passenger miles from 2010 to 2023 with forecasts of selected ARIMA model. The ARIMA model predicts a sharp increase in RPM during the Covid period with narrow prediction intervals, indicating that the forecasted values are high in precision as they are expected to be close to the actual data. However, the actual data shows a sharp decrease during the same period. The difference between the estimated and actual data is due to the unforeseen event of Covid-19 pandemic where movement restrictions were implemented and affected the demand of airline travel.

Comparison with an ETS model

Forecasts for the "Covid" time period using ETS model are produced to examine and compare the performance of both forecasting models in prediction as both models have different assumptions and strengths.

The ETS model computed automatically by R is ETS (M, Ad, M) whereby the model implies that the "Pre-Covid" data have components of multiplicative errors and seasonality with additive damped trend. The ETS model also has an alpha value of 0.352, indicating that 35.2% of the forecasts are based on most recent observations and 64.8% are based on previous observations; beta value of 0.026, indicating that 2.6% of slope are based on most current slope and 97.4% of slope are based on most previous slope; near zero value of gamma, indicating the seasonal component value changes very little over time; and phi value of 0.978, indicating a low damping.

Table 3.2: Parameter estimates for ETS (M, Ad, M) model on "Pre-Covid" data

ETS (M, Ad, M)	
Smoothing Parameters	
Level: Alpha (α)	0.352
Trend: Beta (β)	0.026

Seasonal: Gamma (γ)	1e-04
Damped: Phi (ϕ)	0.978
Initial States	
Level (l)	65,364,573.478
Trend (b)	143,164.8325
Seasonal Components (s) s0, s-1, s-2, s-3, s-4, s-5, s-6, s-7, s-8, s-9, s-10, s-11	0.973 0.915 0.990 0.947 1.121 1.167 1.109 1.045 0.985 1.029 0.821 0.898
Sigma	0.0137
AIC	3945.968
AICc	3952.674
BIC	3996.292

The comparison between ARIMA and ETS models in forecasting RPM during the "Covid" time period is shown as **Figure 3.2** below.

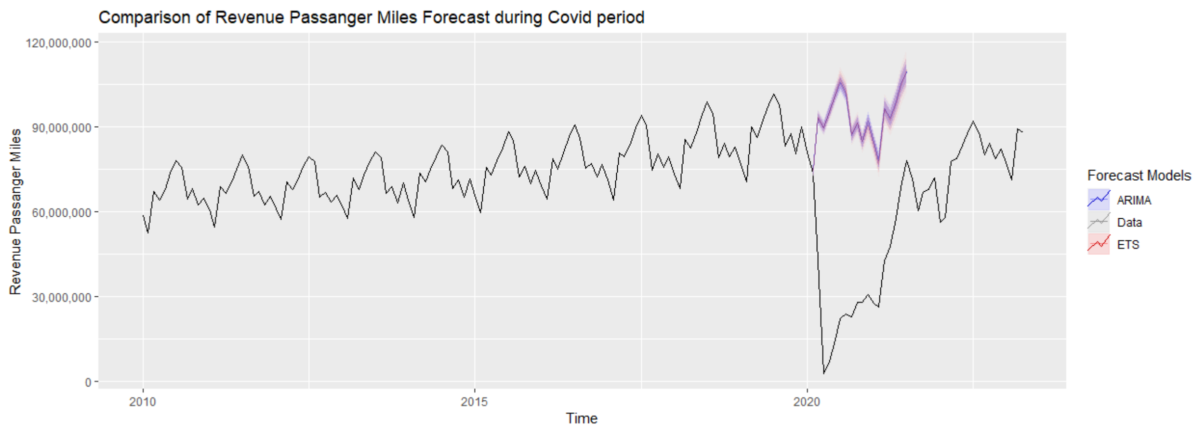

Figure 3.2: Forecast for "Covid" time period using ARIMA and ETS model

Both forecasting models produce similar results with slightly wider prediction intervals from the ETS model. Both models predicted a sharp increase in RPM during the period. We can also see that RPM values peaked around July or August which is the summer season in the United States. The similarity in predicted values by both models suggests that the models have a high level of confidence in predicting. Similar to the graph in **Figure 3.1**, the actual data declined sharply during the period due to the unforeseen event of Covid-19 which peaked in March 2020 as it was the month when travel restrictions were implemented.

We also evaluate the performance of both forecasting models over the test set. The results are shown in **Table 3.3** below. Overall the accuracy for both models is very similar which reflects the similar results in **Figure 3.2**. However, the performance of the ETS model is slightly more accurate than the ARIMA model based on RMSE, MAE, MPE, MAPE, and MASE as the ETS model has lower values of these performance metrics.

Table 3.3: Comparison of out-of-sample forecast accuracy of ARIMA and ETS models

Test set	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**
ARIMA $(0,1,1)(0,1,2)_{12}$	61,768,733	57,717,793.9	-407.736	407.736	23.411
ETS (M, Ad, M)	61,645,325.4	57,482,357.2	-406.824	406.824	23.316

Covid-19 pandemic has brought a huge impact on the passenger air transport industry where they face huge financial loss due to the severe demand decline. To aid decision-making for airline companies, we calculated the forecasted loss in US RPM by subtracting the total of actual RPM values from the total of forecasted RPM values from ARIMA and ETS models shown in **Table 3.4**. The forecasted loss in RPM from the ARIMA model is 1,038,920,289 miles whereas the forecasted loss in RPM from the ETS model is 1,034,682,429 miles, which is slightly smaller. This may be due to the fact that the ETS model's forecast accuracy is slightly better shown in **Table 3.3**.

Table 3.4: Calculation of Forecasted Loss in RPM from ARIMA and ETS model

	ARIMA $(0,1,1)(0,1,2)_{12}$	**ETS (M, Ad, M)**
Total Forecasted RPM	1,684,916,047	1,680,678,187
Total Actual RPM	645,995,758	645,995,758
Forecasted Loss (Total Forecasted RPM - Total Actual RPM)	1,038,920,289	1,034,682,429

Conclusion

Covid-19 pandemic has brought a huge impact on the passenger aviation industry due to travel restrictions and social distancing measures implemented. The results from ARIMA and ETS models have proven that forecasting models could not capture unprecedented events such as the Covid-19 pandemic. However, through forecasting RPM, we can find out the forecasted loss by the aviation industry.

Through analysing and transforming the original data into stationary data that is suitable for the ARIMA model, we identified all possible combinations of ARIMA models based on the ACF and PACF plots. Then, examination and testing of potential ARIMA models are done to identify the best ARIMA model based on AICc after modifications on the models. After identifying the best ARIMA model which is ARIMA $(0,1,1)(0,1,2)_{12}$, forecasts using the model based on the full dataset are produced and plotted against the actual data along with a forecast from the ETS (M, Ad, M) model to compare the performance of both forecasting models.

Here, we found out that both forecasting models gave slightly different results for the forecasted loss in RPM as the ARIMA model forecasted a loss of 1,038,920,289 miles while the ETS model forecasted a loss of 1,034,682,429 miles which is slightly lower than the ARIMA model. The almost identical results from both models suggest a high level of confidence in the predictions while the slight difference from the results proved that forecasting models have different methodologies and strengths on different data characteristics. With the knowledge of forecasted loss in RPM, aviation companies can make informed planning and decision-making for better adaptation and innovation such as improving their sales by increasing their aircraft hygiene and safety to gain their customers' trust. They need to come up with strategies to mitigate any future global crisis like this.

Reference

Naseer, S., Khalid, S., Parveen, S., Abbass, K., Song, H., & Achim, M. V. (2023). COVID-19 outbreak: Impact on global economy. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.1009393

Salas, E. B. (2022, September 13). Coronavirus: impact on the aviation industry worldwide – statistics & facts. Statista. Retrieved September 15, 2022 https://www.statista.com/topics/6178/coronavirus-impact-on-theaviation-industryworldwide/

Warnock-Smith, D., Graham, A., O'Connell, J. F., & Efthymiou, M. (2021, July). Impact of COVID-19 on air transport passenger markets: Examining evidence from the Chinese market. Journal of Air Transport Management, 94, 102085. https://doi.org/10.1016/j.jairtraman.2021.102085