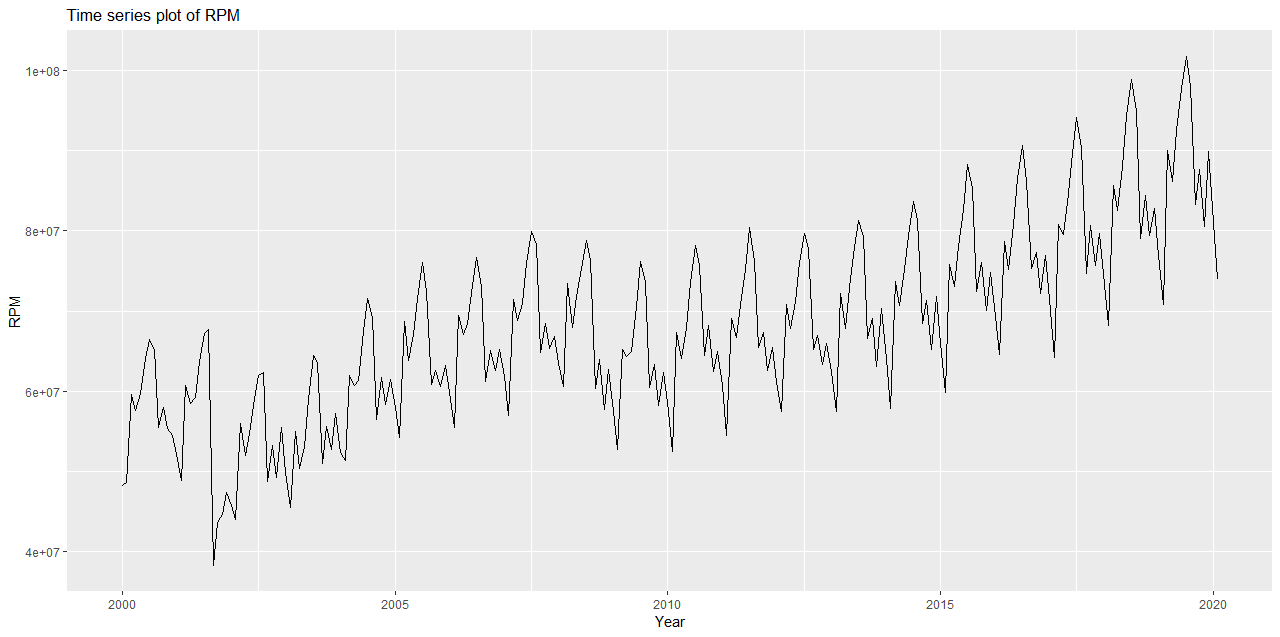
FIN642 - Revenue Passenger Miles Project Report

Group 6

Yusi Luo, Hao Deng, Shuai Qin

**1. Introduction**

For this project, we used a dataset called Revenue Passenger Miles (RPM), which calculated by multiplying the number of paying passengers by the distance travelled. RPM is a common airline statistic that shows the number of miles travelled by paying passengers and is widely used by airline companies and public agencies to access airline business performance. The RPM data spans from Jan 2000 to Feb 2020 and contains 242 observations. An initial plot or RPM is as follows:





图片包含 游戏机

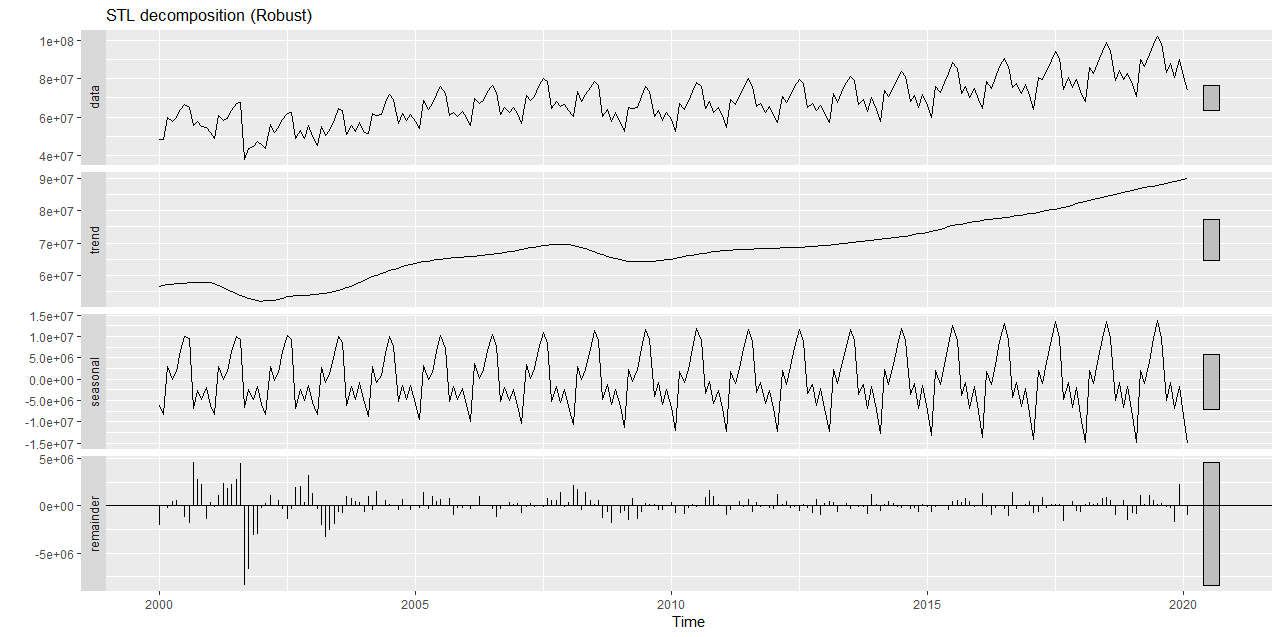
描述已自动生成

As we can identify repetitive patterns over the time horizon, the time series plot demonstrates strong evidence for seasonal effect. Also, the time series has an increase trend on the whole.

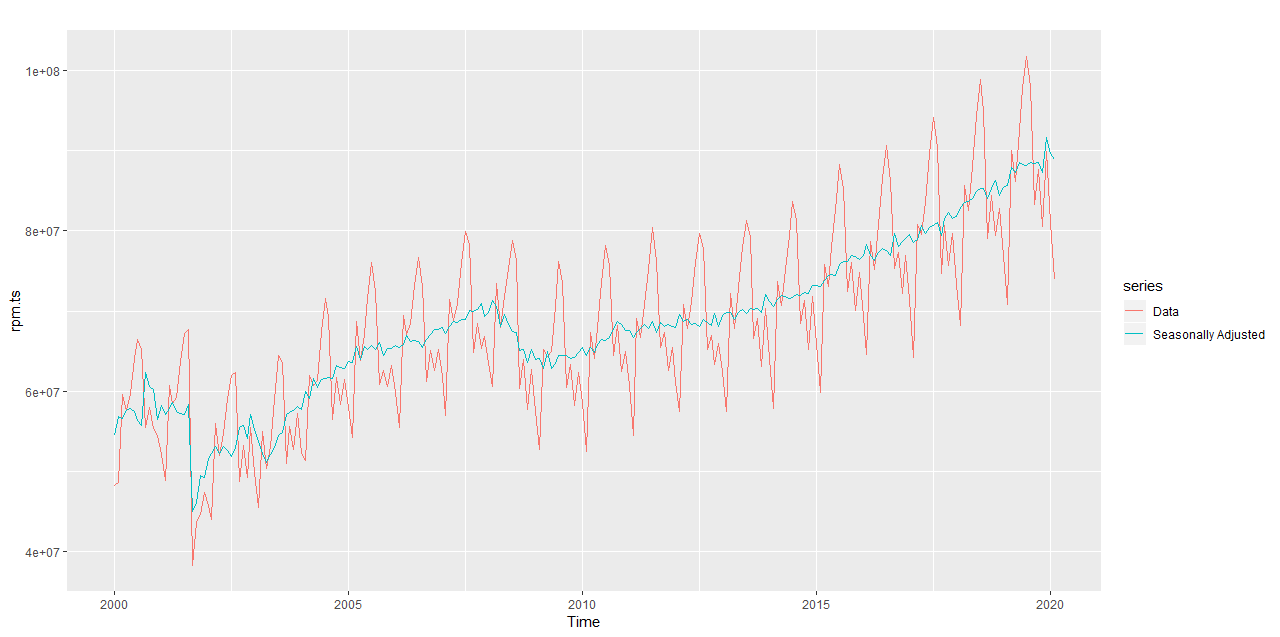
We can see that the maximum value of revenue passenger miles is 101,794,184 in July, and the minimum is 38,207,236 in February. July and August have most revenue passenger miles because this is summer vacation. And for February, this is the time to work and study, so there are few people to take airplane.

**2. Model selection**

Our analysis begins from the decomposition models. We apply the robust STL decomposition and obtained the following results:



As expected, the evidence does suggest a significant trend and seasonality within our data. The seasonal effect is further verified as follows:



There is also a time trend which indicates the increasing RPM over the past ten years. This is expected, as the tourism industry has been expanding thanks to the internet and social media that promoted travel opportunities globally. The advance in airplane technology and increasing competition in airline industry would also make air travel more affordable, thereby increasing the total usage of air travel.

The first model is **Seasonal Naïve**. Residuals are correlated as shown by both the LB test and the ACF and do not appear to be normal (they have a long-left tail). There is considerable information remaining in the residuals which has not been captured with the seasonal naive method.

图片包含 游戏机

描述已自动生成

截图里有图片

描述已自动生成手机屏幕截图

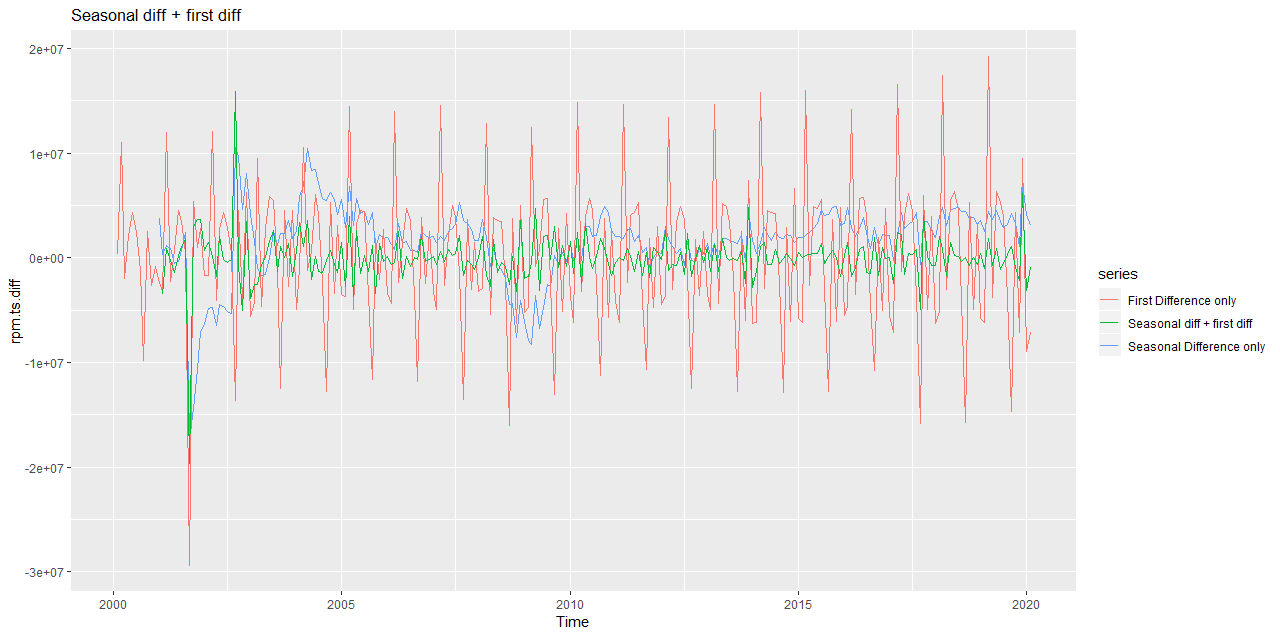
描述已自动生成

Moving to the next section, we use **Holt-Winters models**. The multiplicative method is preferred when the seasonal variations are changing proportional to the level of the series. And we can see that the data has strong seasonal effect, so Holt-Winter is very appropriate method. We will also compare another method that Holt-Winters method with a damped trend and multiplicative seasonality which can often provide accurate and robust forecasts for seasonal data. We can see from the picture that Holt-Winters’ multiplicative method is better than Holt-Winters’ damped method. Because the RMSE is lower.

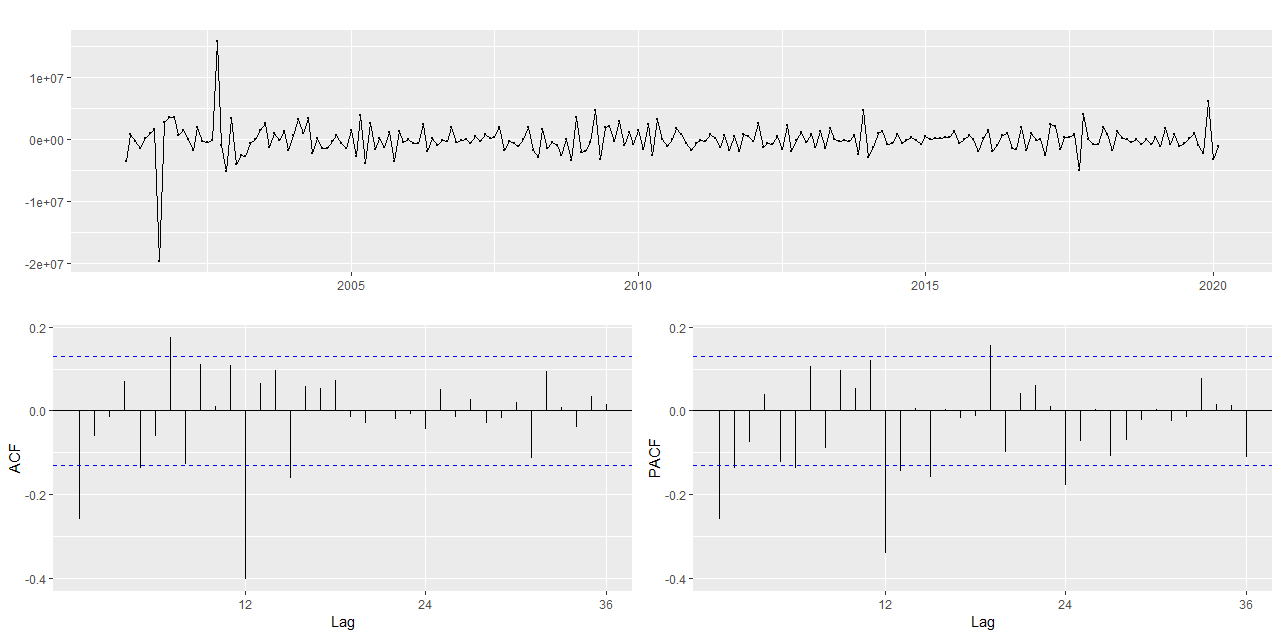
手机屏幕截图

描述已自动生成

Finally, we carry out analysis with **ARIMA** and **SARIMA** models. We begin by checking the number of differences it takes to stabilise the stationarity of the data. The red line represents first differenced data, which removes the trend but is clearly seasonal and non-stationary. The blue line represents 12-times differenced data, which does not contain any seasonality but contains some trends and cycles. The data with both first difference and 12-times difference, the green line, looks approximately stationary.



Now that we are certain to take d=1 and D=1 for both non-seasonal component and seasonal component, our model would look like ARIMA(?,1,?) X (?,1,?)12, where we need to determine the MA and AR compotents. We use correlograms to examine the AR and MA behaviour:



The spike at beginning of ACF and PACF suggest that the model may contain MA (1) component, which is supported by our previous SEATS results. The reducing spikes at 12 and 24 lag on PACF indicates strong basis for seaonsal MA (1) effect. As such, we would compare three different models: model 1 is ARIMA(1,1,1) X (0,1,1)12, model 2 is ARIMA(0,1,1) X (1,1,1)12, and model 3 is ARIMA(1,1,1) X (0,1,2)12. The three models are fitted and their metrics are extracted:



The lowest AICc is for model2 ARIMA(0,1,1)(1,1,1), therefore it is the preferred model for forecasting. We further check the residuals of model 2:

图片包含 游戏机, 大, 水

描述已自动生成手机屏幕截图

描述已自动生成

Unfortunately, model 2 also has an outlier in its residual, however, it passes the portmanteau test with a large p-value. The forecast of model 2 is visualised as follows:

图片包含 游戏机

描述已自动生成

Finally, we would like to compare the three classes of models with their performances over test set. If a model performs well on test set, then we would be confident that it will perform well in real life forecast. The forecasts made by three models are as follows:

图片包含 游戏机

描述已自动生成

With visual inspection, the Holt-winter’s multiplicative model is closest to the actual data, then followed by ARIMA model and Seasonal Naïve model. This is also evidenced by their accuracy scores:

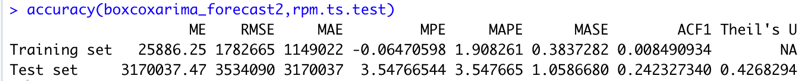
手机屏幕截图

描述已自动生成

As we mention above, we did not perform boxcox transformation because we consider the differenced trend to be relatively stationary. To compare the final performance of ARIMA model, we then apply boxcox transformation to our dataset. The result was not optimistic compare to the original model results. ARIMA model **without** boxcox transformation has better performance over ARIMA model with boxcox transformation.

A screenshot of a cell phone

Description automatically generated



In conclusion: **The Holt-Winter multiplicative model** has lowest RMSE, defeating other two models.