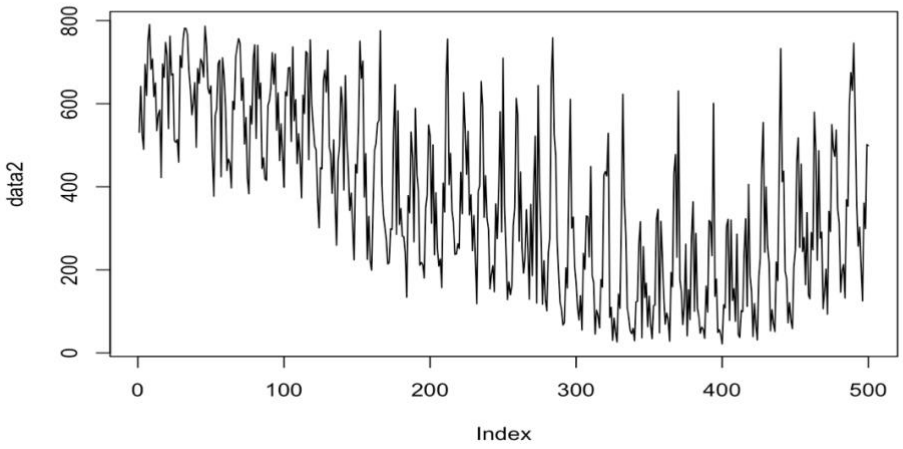
Analysis of Data 2 **Man Chong Chan(3033281903)**

Exploratory Data Analysis:

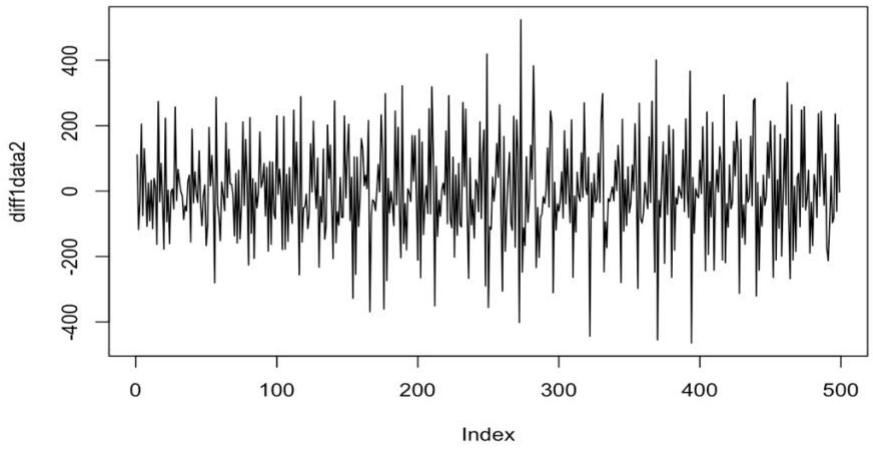
The time series plot of data 2 is shown below in Fig 1. As it is shown in Fig 1, the time series seems to have a stable variance over time. However, it is obvious that it has a downward trend from data point 1 to 350 and then an upward trend from 350 to 500.

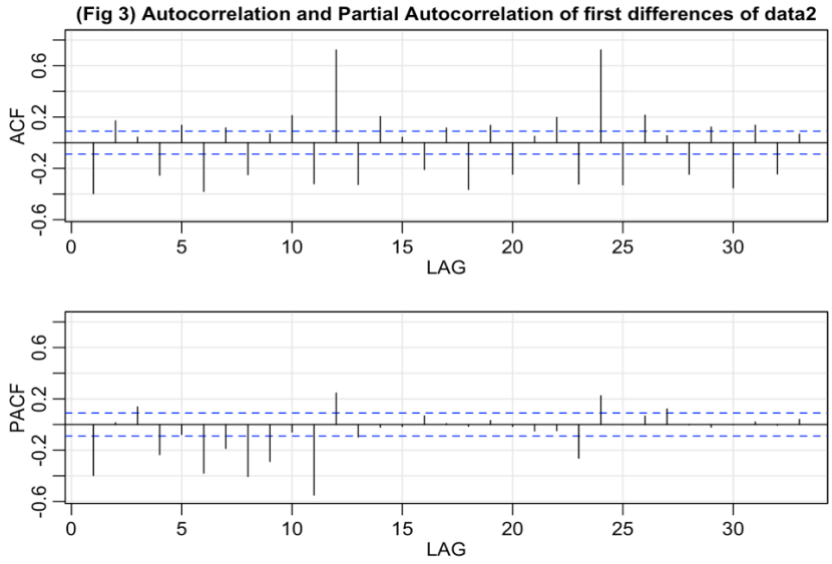
(Fig 1) Time series plot of data2



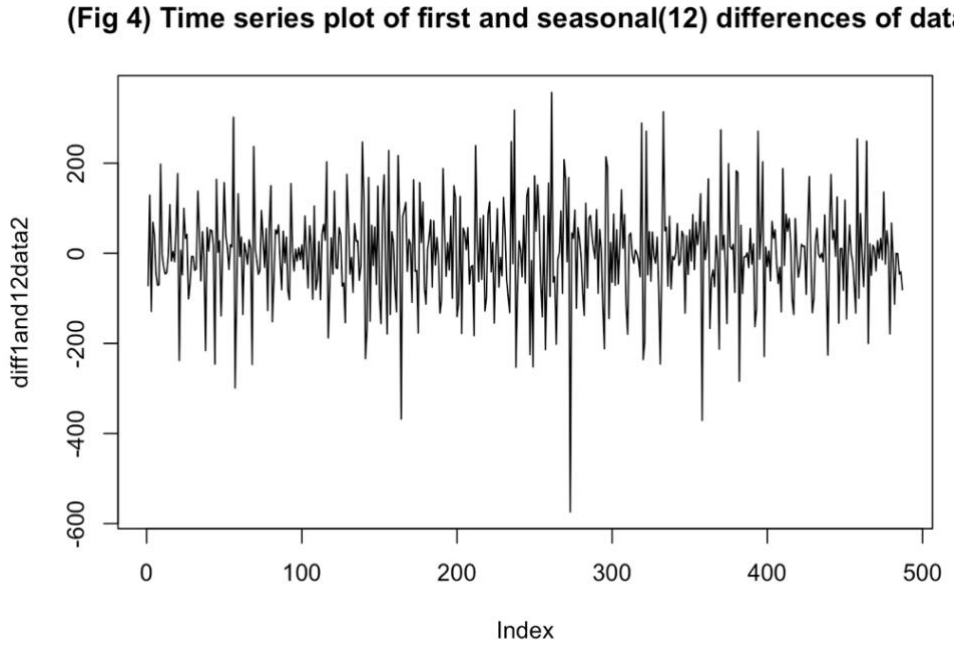
To deal with the trend, first order non-seasonal difference was performed and consequently the time series plot look a lot more stationary than before. (Fig 2) Its ACF and PACF are also shown below. (Fig 3)

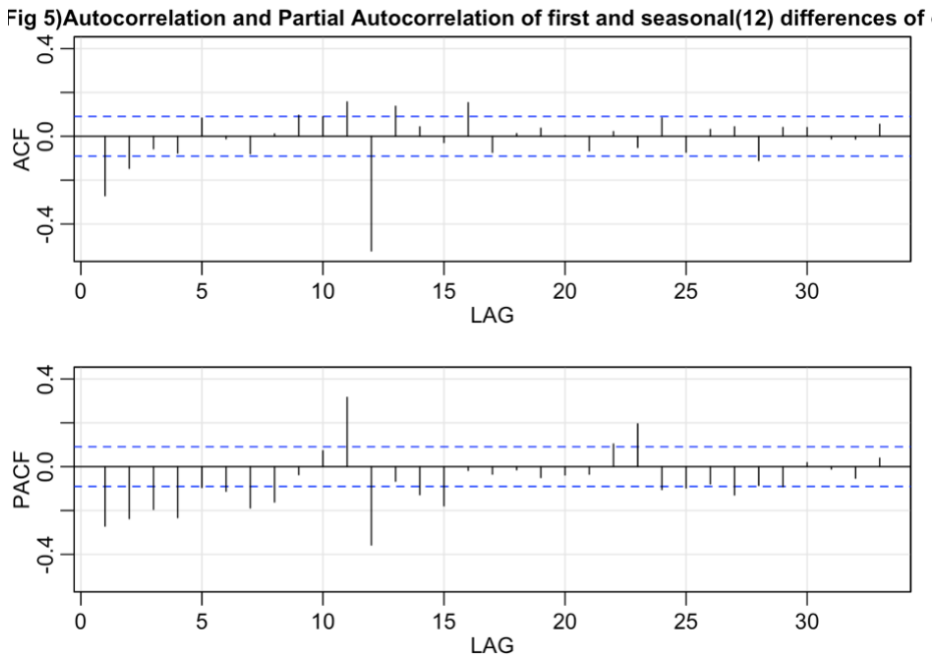
(Fig 2) Time series plot of first differences of data2





As we can see in Fig 2, the trends are removed after the differencing. However, there exist a strong seasonality. We can confirm this by looking at the ACF of first differences of data 2 (Fig 3), there a is large positive spike at every lag of multiples of 12. To eliminate the seasonality, seasonal differencing was applied to the data. The time series plot after first and seasonal difference is shown below in Fig 4, and The ACF and PACF are shown below in Fig 5.



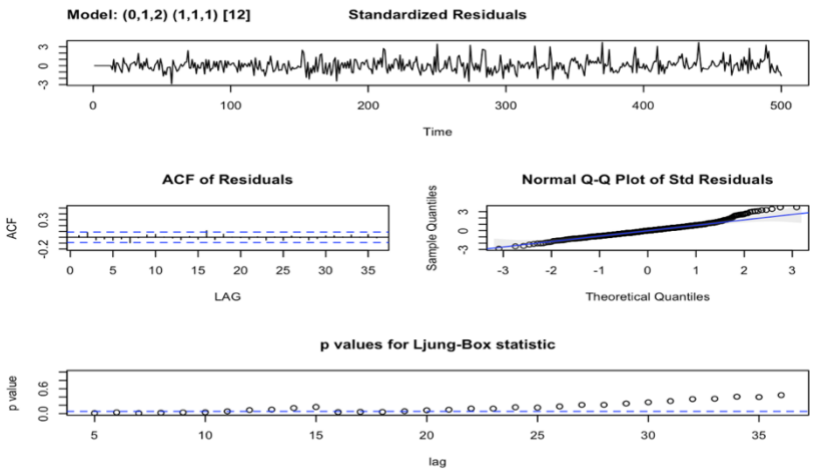


Finally, after one differencing and one seasonal differencing, the time series looks stationary with zero mean and constant variance.

Identifying suitable ARMA model:

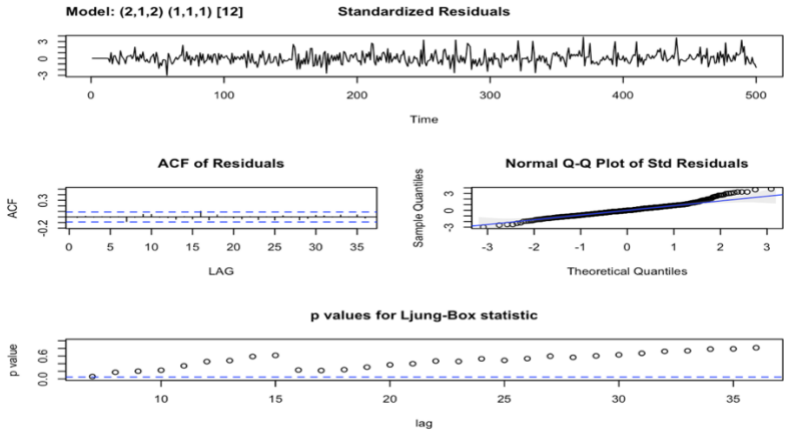
In the ACF (Fig 5), there is a huge negative autocorrelation at lag 12 (a year), and then a bunch of small autocorrelations between lag 1 and lag 12; a few small autocorrelations after lag 12. Meanwhile in the PACF (Fig 5), there are bunch of big autocorrelations from lag 1 to lag 12 (a year), and some relatively small autocorrelations after lag 12. It seems like both of the ACF and PACF cut off after lag 12 (a year). This suggests the model we want should contain SARMA(1,1) as its seasonal component.

Next, we look at the contribution from the ARMA part to ACF/PACF (Fig 5), which are the lags between 1 and 12 (lags within a year). One possible interpretation is the ACF cuts off sharply at lag 2 and the PACF tapers off, so MA(2) might be reasonable for describing the non-seasonal ARMA part. Thus, We try to fit ARIMA(0,1,2)(1,1,1)[12] and call this our model 1.



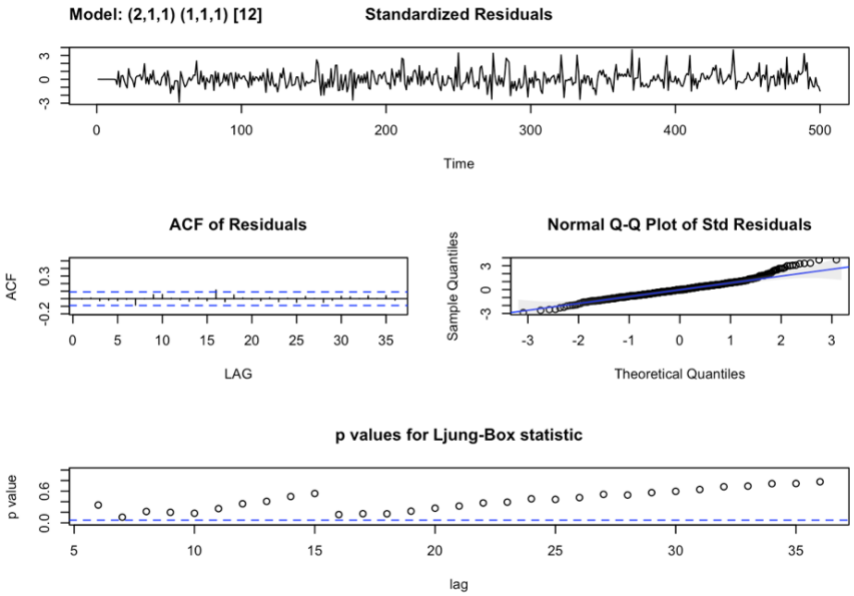
Looking at the sarima() fitting output of model 1, the Ljung-Box statistics are not too good : the second half of it are insignificant but the first half are significant. As a result, we need to reexamine the ACF/PACF and would not take model 1 into consideration.

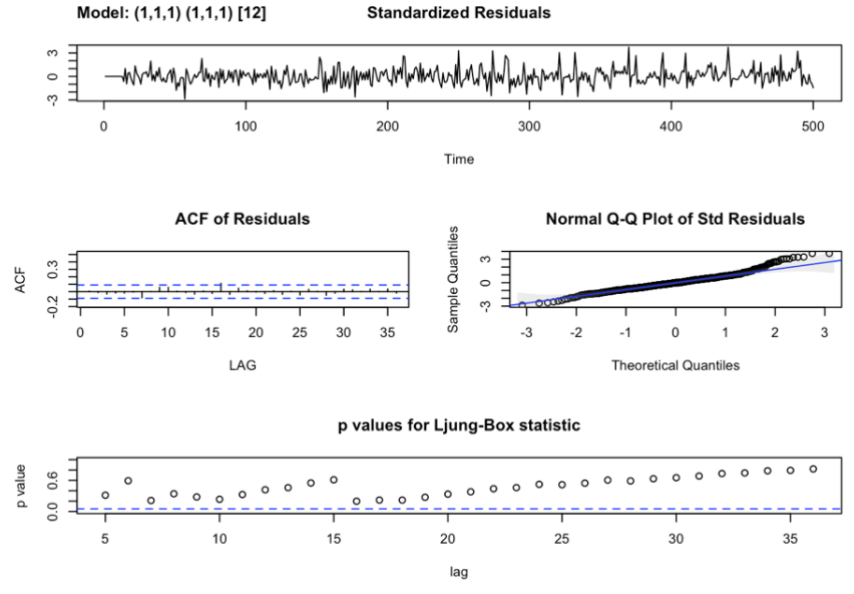
Another possible interpretation is that both ACF and PACF tapers off. Notice that the ACF and PACF start cutting off at lag 2 and lag 8 respectively. Since the parameter q cannot exceed the maximum lag cutting off in ACF and neither can p exceed the maximum lag cutting off in PACF, the non-seasonal parameter p and q should be under 8 and 2 respectively. Considering ACF cuts off rather sharply at lag 2, the AR order p should be a small value and also, we do not want our p as large as 8, which may lead to overfitting. As a result, we try to fit ARIMA(2,1,2)(1,1,1)[12] and call this our model 2.



Looking at the sarima() fitting output of model2, Ljung-Box statistics looks really good: most of them are all out side of the band. The standardized residuals look like an iid mean 0, variance 1 sequence according to the ACF of residuals.

Since we concluded that non-seasonal q should be under 2 and we do not want p to be too large, we can also try ARIMA(2,1,1)(1,1,1)[12] and call this our model 3 and ARIMA(1,1,1)(1,1,1)[12] and call this our model 4.





Looking at the sarima() fitting output of model 3, Ljung-Box statistics also looks even better than model 2: they are all out side of the band. The standardized residuals also look like an iid mean 0, variance 1 sequence according to the ACF of residuals.

Same as model 3, the sarima() fitting output of model 4 seems really good as well. The Ljung-Box statistics are all insignificant. The standardized residuals look like an iid mean 0, variance 1 sequence according to the ACF of residuals.

Right now, there are 3 models that could be a good fit of data 2 and we will choose one out of these by comparing their AIC, AICc, BIC and also cross validation value(CV). The values which are collected using R are shown in the table below as a summary.

MODELS	AIC	AICC	BIC	CV
(MODEL2)ARIMA(2,1,2)(1,1,1)[12]	9.558734	9.56319	8.60931	77,177
(MODEL3)ARIMA(2,1,1)(1,1,1)[12]	9.559429	9.56377	8.60158	75,497
(MODEL4)ARIMA(1,1,1)(1,1,1)[12]	9.556032	9.56028	8.58975	75,529

According to the table, model 4 ARIMA(1,1,1)(1,1,1)[12] might be the best fit among because it has the best AIC, AICc, BIC and second best CV. Thus, I would choose model 4 as my model.

Model fitting and forecasting:

Using sarima() in R to get estimated Parameters:

AR1 = 0.3655, MA1 = -0.9275, SAR1 = -0.0205, SMA1 = -0.8529

Using predict() in R to get 10 Forecasts:

380.2567, 583.4951, 377.6791, 345.1008, 226.6389, 291.3197, 246.5470, 191.7618, 367.7683, 342.1272