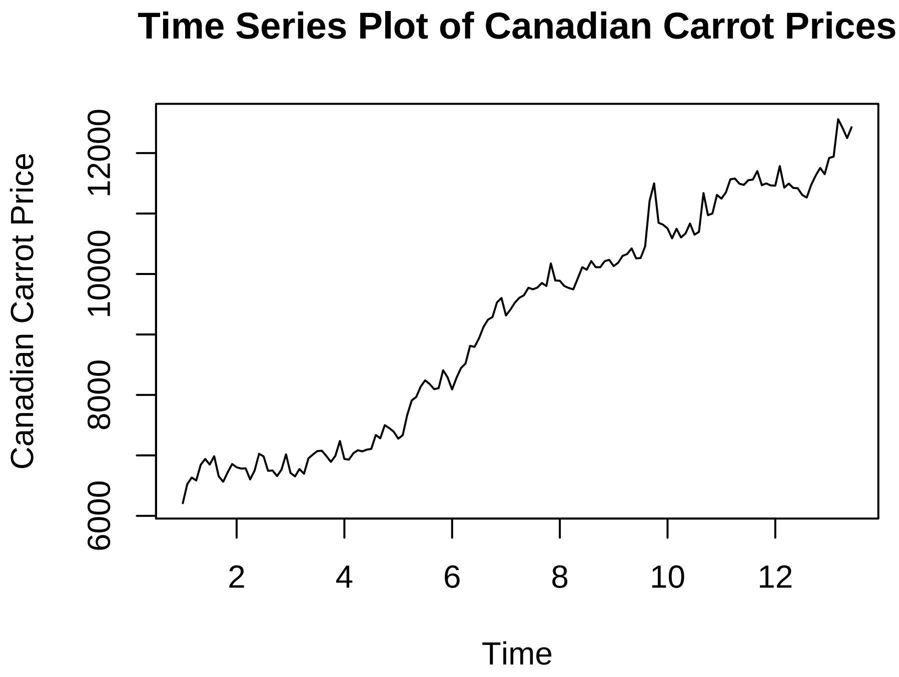
A Time Series Analysis on Forecasting Canadian Carrot Prices using SARIMA model

Intro

This report specifically deals with the forecasting and prediction of Canadian carrot prices for the next twelve observations (i.e., predicting the carrot prices for the next 12 months) using various time series models. Different models have been fitted to the collected data and each model's limitations and weaknesses will be further evaluated and discussed.

Firstly, to get a rough idea of the observations/data we have already collected, a visual representation of this time series plot is generated as follows:

```r
data <- read.csv('Canadian Carrot Prices.csv', header=FALSE)$V2 #get second col only
remove_header = data[-1] #remove header "Canadian Carrot Prices"
numeric_data = as.numeric(levels(remove_header))[remove_header] #convert to numeric data
num_data = length(numeric_data) #150 total observations
ts = ts(numeric_data, freq=12) #convert to time series
plot(ts, ylab = "Canadian Carrot Price", main = "Time Series Plot of Canadian Carrot Prices")
```



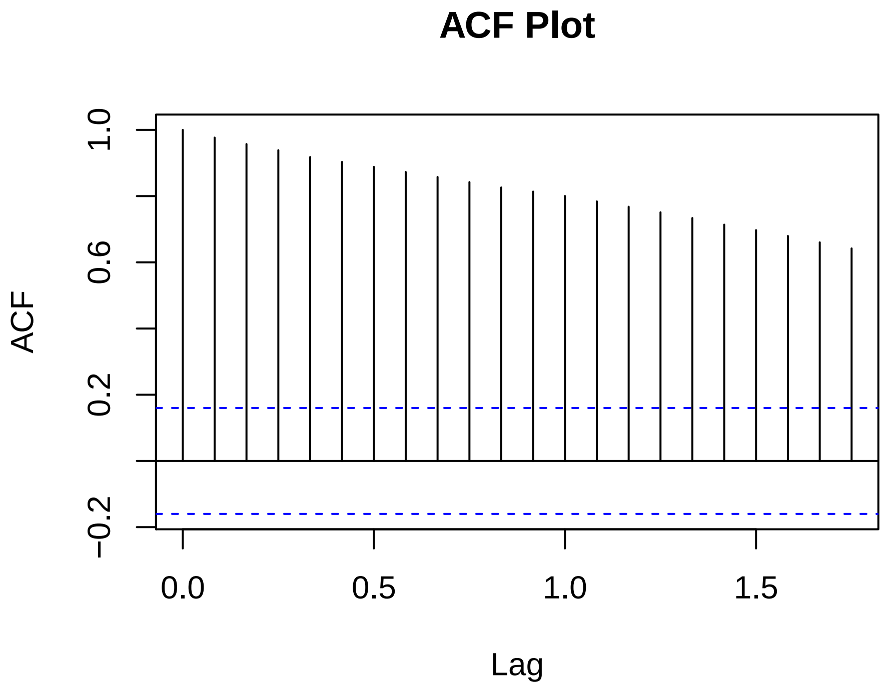
(Note that this plot is based on 150 observations).

First, we can already observe from the original time series plot (without applying any transformations) that there is a clear increasing trend with some evidence of seasonality, indicated by the rise and falls of the "triangular-like peaks" presented in a regular interval. This also make sense as it suggests that the carrot prices might have been strongly influenced by different time of the year (i.e., prices might be increased/decreased during different seasons). Hence, in order to further investigate these problem, the ACF plot is generated as follows:

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
acf(ts, main = "ACF Plot")
```

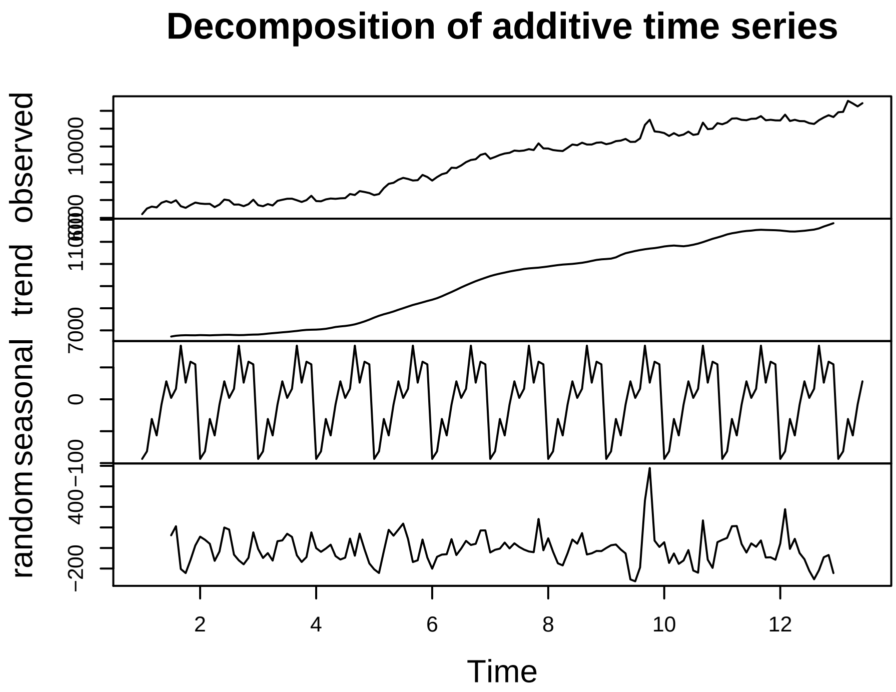


```
adf.test(ts)
```

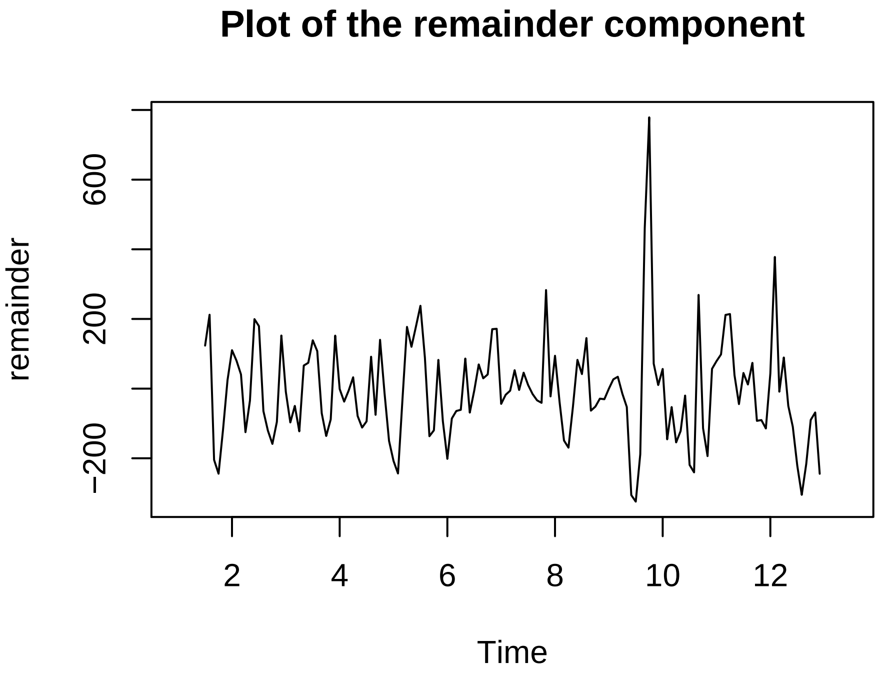
```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts
## Dickey-Fuller = -2.5578, Lag order = 5, p-value = 0.344
## alternative hypothesis: stationary
```

As we can see from plot above, the auto-correlation is always very high and it is also decreasing very slowly linearly, which is similar to that of a Random Walk Process, indicating that this series is not stationary. The Augmented Dickey-Fuller Test (ADF test) also shows a high p-value of 0.344, which fails to reject that the series is non-stationary at the 95% confidence level. This non-stationarity is also reflected by the plot of the original series where we can see that carrot prices are not fluctating around a centered mean with a constant variance. This all indicates further preprocessing is required. I will first perform a classical decomposition as follows:

```
dcmp = decompose(ts)
plot(dcmp)
```

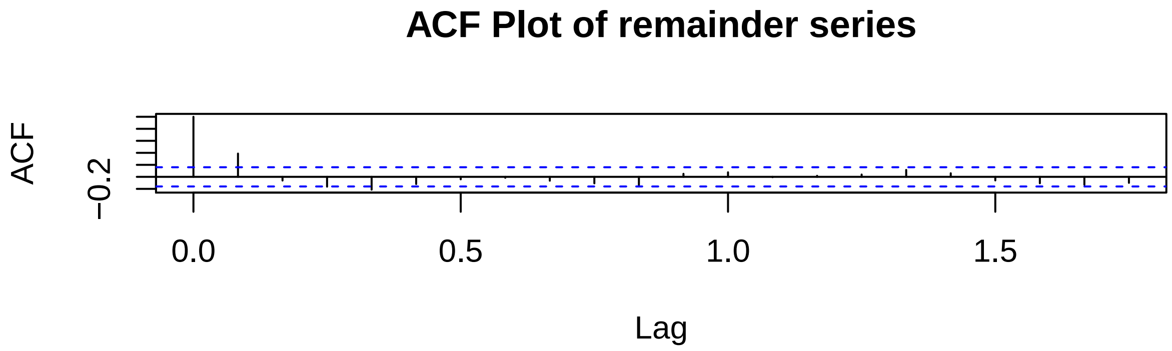


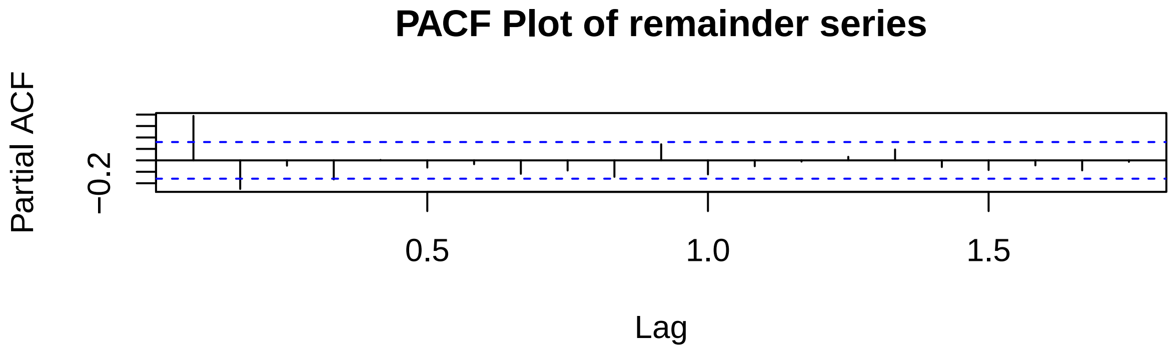
```
remainder = dcmp$random
plot(remainder, main = "Plot of the remainder component")
```



Based on the classical decomposition, we can conclude that there is indeed somewhat an increasing trend and a strong seasonal pattern. However, the `remainder` (or random) component of the classical decomposition seems somewhat stationary as the mean of the series is now centered and fluctuating accordingly with a more or less consistent variance with some outliers around the 9th year (as I have removed the `trend` and `seasonal` components of the original series). Now, I will attempt to determine the model parameters using PACF and ACF plots, as well as taking suggestion from using the auto.arima() function on this `remainder` series.

```r
library(astsa)
library(forecast)
par(mfrow=c(2,1))
acf(remainder, na.action= na.pass, main= "ACF Plot of remainder series")
pacf(remainder, na.action= na.pass, main= "PACF Plot of remainder series")
```


ACF Plot of remainder series


PACF Plot of remainder series

```r
auto_remainder = auto.arima(remainder) #auto.arima() on remainder series
auto_remainder
```

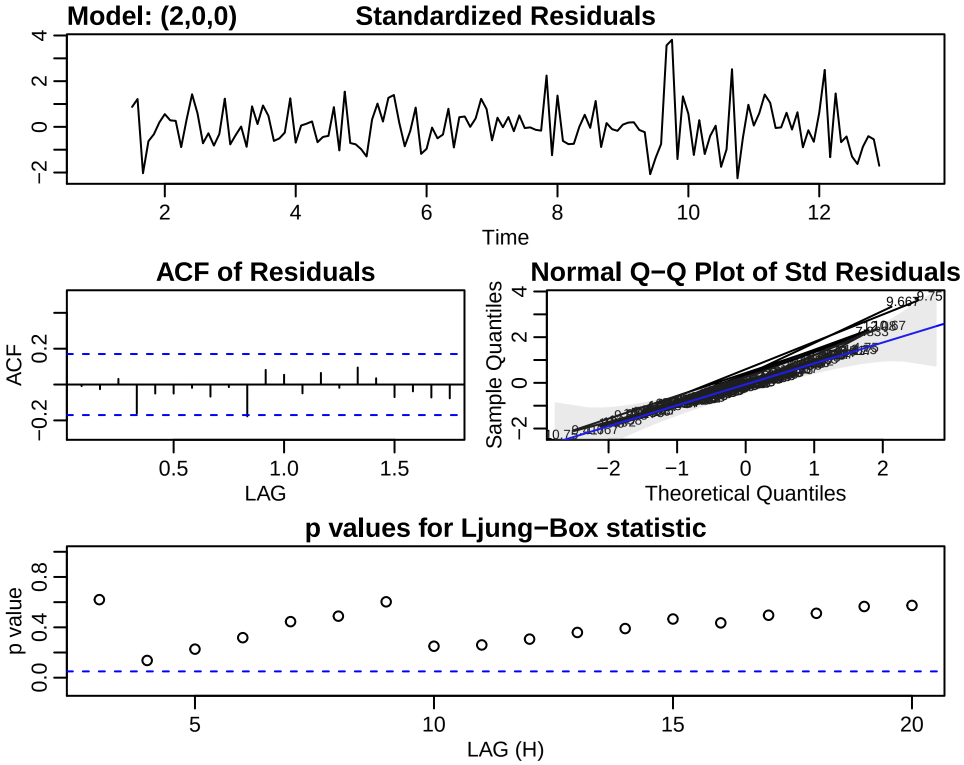
```
## Series: remainder
## ARIMA(0,0,1) with zero mean
##
## Coefficients:
##          ma1
##       0.5157
## s.e.  0.0780
##
## sigma^2 estimated as 18179:  log likelihood=-872.22
## AIC=1748.44   AICc=1748.53   BIC=1754.29
```

As we can see from the plots above, the ACF cuts off at lag 1 suggesting a `MA(1)` model and the PACF also kind of cuts off at lag 2, suggesting an `AR(2)` model. The `auto.arima()` function on the remainder series also suggests a MA(1) model. Now, we will use `astsa::sarima()` function to compare the diagnositics plots of these 2 models to decide on the better model.

Diagnositics for AR(2) model:

```
library(astsa)
AR_2_model = arima(remainder, order= c(2,0,0))
sarima(remainder, 2,0,0) #AR(2) for diagnostics okit
```

```
## initial  value 5.013221
## iter   2 value 4.916860
## iter   3 value 4.908930
## iter   4 value 4.897637
## iter   5 value 4.897483
## iter   6 value 4.897476
## iter   6 value 4.897476
## iter   6 value 4.897476
## final  value 4.897476
## converged
```



Model: (2,0,0) Standardized Residuals

ACF of Residuals

Normal Q−Q Plot of Std Residuals

p values for Ljung−Box statistic

```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = xmean, include.mean = FALSE, transform.pars = trans,
##     fixed = fixed, optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1      ar2    xmean
##       0.4924  -0.2539  -7.9231
## s.e.  0.0828   0.0831  14.9606
##
```

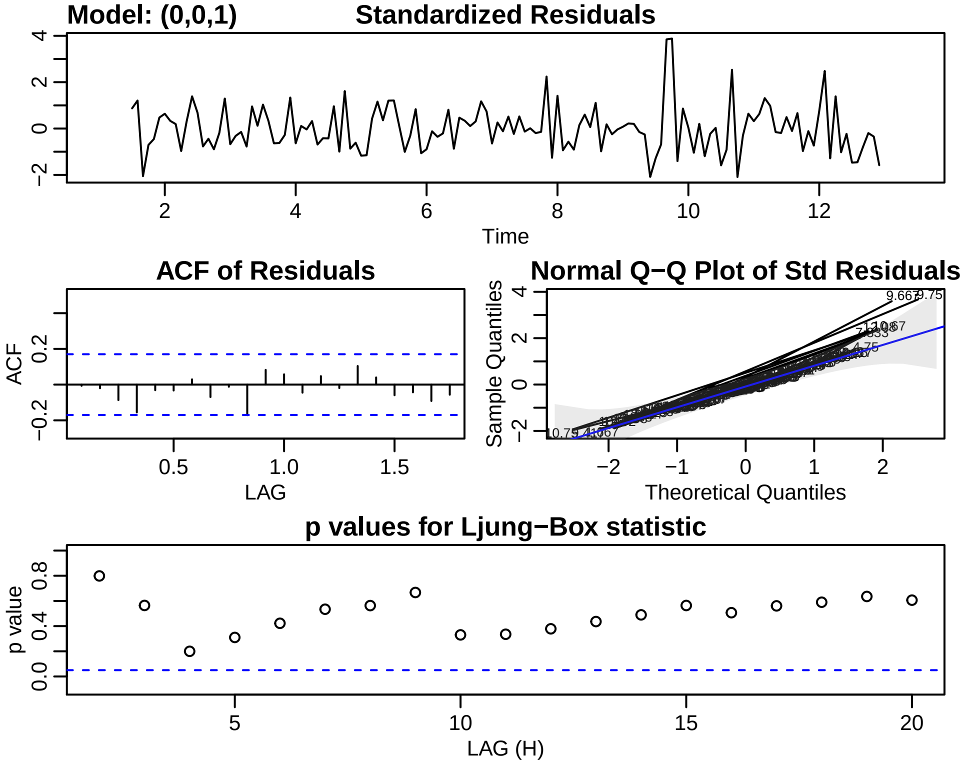
```
## sigma^2 estimated as 17904:  log likelihood = -871.67,  aic = 1751.33
##
## $degrees_of_freedom
## [1] 135
##
## $ttable
##       Estimate      SE t.value p.value
## ar1     0.4924  0.0828  5.9458  0.0000
## ar2    -0.2539  0.0831 -3.0546  0.0027
## xmean  -7.9231 14.9606 -0.5296  0.5973
##
## $AIC
## [1] 11.67554
##
## $AICc
## [1] 11.67663
##
## $BIC
## [1] 11.7536
```

First, we can see from the ACF plot of the residuals that they are all within the 95% confidence intervals, indicating that there is no correlation between the residuals (suggesting a good fit of the model). The Normal Q-Q plot suggests that there are a few outliers (on both end of the tails), making the normality of the residuals to be slightly violated (but the bulk of the residuals are still following a Normal distribution). All p-values of the Ljung-Box test are all above the 5% confidence level (blue dashed line), indicating that the model has no serial correlation with 95% confidence level. Finally, the AIC value for the AR(2) model is 11.676, AICc is 11.677, and BIC is 11.753 and we will compare these values with our next potential model, the MA(1) model.

Diagnositics for MA(1) model:

```
MA_1_model = arima(remainder, order= c(0,0,1))
sarima(remainder, 0,0,1) #MA(1)
```

```
## initial  value 5.013221
## iter   2 value 4.909319
## iter   3 value 4.903417
## iter   4 value 4.900784
## iter   5 value 4.900699
## iter   6 value 4.900698
## iter   7 value 4.900698
## iter   7 value 4.900698
## iter   7 value 4.900698
## final  value 4.900698
## converged
```



```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = xmean, include.mean = FALSE, transform.pars = trans,
##     fixed = fixed, optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ma1    xmean
##       0.5152  -8.1218
## s.e.  0.0780  17.2728
##
```

```
## sigma^2 estimated as 18019:  log likelihood = -872.11,  aic = 1750.22
##
## $degrees_of_freedom
## [1] 136
##
## $ttable
##       Estimate      SE t.value p.value
## ma1     0.5152  0.0780  6.6037   0.000
## xmean  -8.1218 17.2728 -0.4702   0.639
##
## $AIC
## [1] 11.66813
##
## $AICc
## [1] 11.66868
##
## $BIC
## [1] 11.72668
```

Now, we can see from the ACF plot of the residuals for the MA(1) model that they are all within the 95% confidence intervals too, indicating that there is no correlation between the residuals (suggesting a good fit of the model). The Normal Q-Q plot suggests that there are a few extreme outliers like the AR(2) model (on both end of the tails), making the normality of the residuals to be slightly violated (but the bulk of the residuals are still following a Normal distribution). All p-values of the Ljung-Box test are all above the 5% confidence level (blue dashed line) as well, indicating that the model has no serial correlation with a 95% confidence level. Finally, the AIC value for the MA(1) model is 11.668, AICc is 11.669, and BIC is 11.727, which is all better than that of the AR(2) model, making this model a more preferable choice out of the two.

Model Adequacy Test of residuals for the two models:

```
Box.test(MA_1_model$resid, lag = 24, type = c("Ljung-Box"), fitdf = 8)$p.value
```

```
## [1] 0.1635133
```

```
Box.test(AR_2_model$resid, lag = 24, type = c("Ljung-Box"), fitdf = 8)$p.value
```

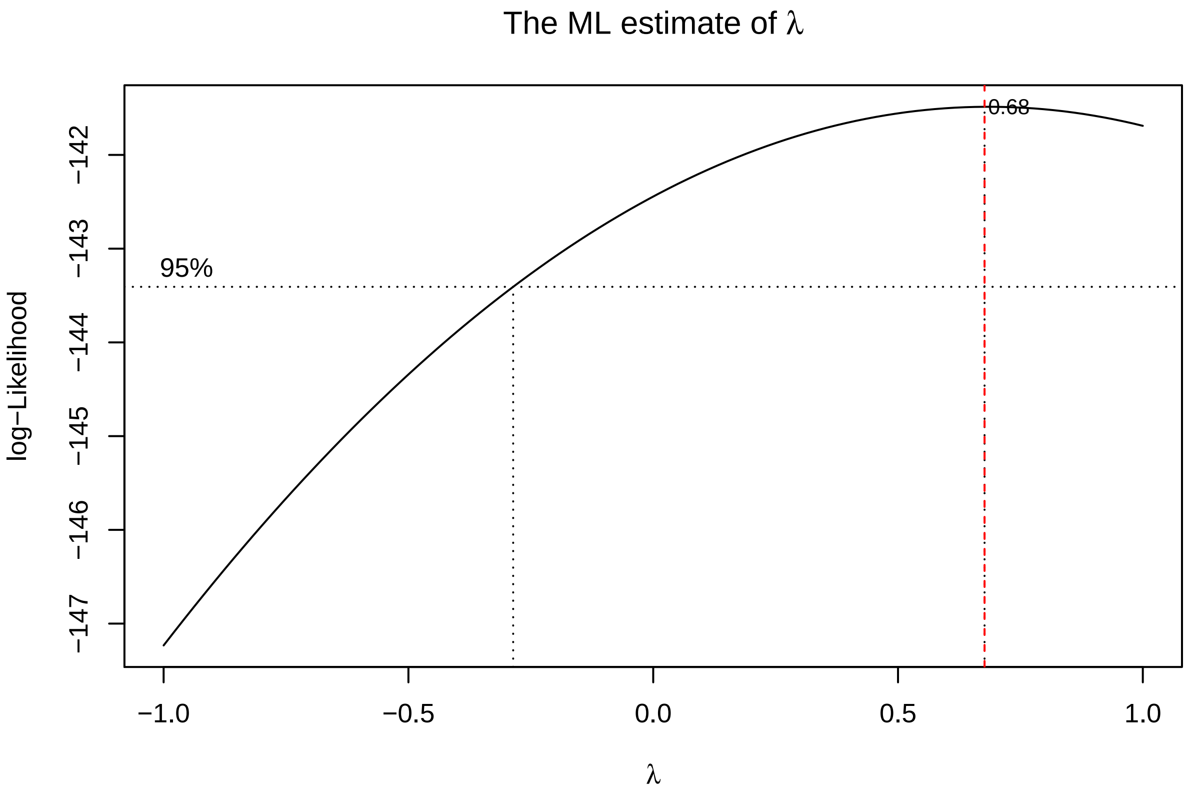
```
## [1] 0.1811292
```

We can see that the p-values of both models passed the (Ljung-Box) model adequacy test at the 95% confidence level, which is consistent to what we have concluded by visual inspection of the diagnositic plots. In other words, we failed to reject the null-hypothesis that the data (residuals) are independently distributed, i.e., we have enough evidence to conclude that there is no serial correlations (of the residuals) for both models.

Further Analysis

However, rather than simply using the classical decomposition, I will also attempt to do some transformations and/or differencing on the original time series directly to deal with non-stationarity directly. A Box-Cox transformation (Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. JRSS B 26 211-246.) to derive an appropriate λ value is done as follows:

```
library(MASS)
par(cex=0.8)
box_cox_transformation<-boxcox(ts~1, lambda=seq(-1,1,0.1))
y_values<-as.numeric(box_cox_transformation$y)
lambda<-box_cox_transformation$x[which.max(y_values)]
abline(v=lambda, col=2, lty="dashed")
text(lambda+0.05,max(box_cox_transformation$y), round(lambda,2), cex=0.85)
title(expression(paste("The ML estimate of ", lambda )), cex=0.85)
```



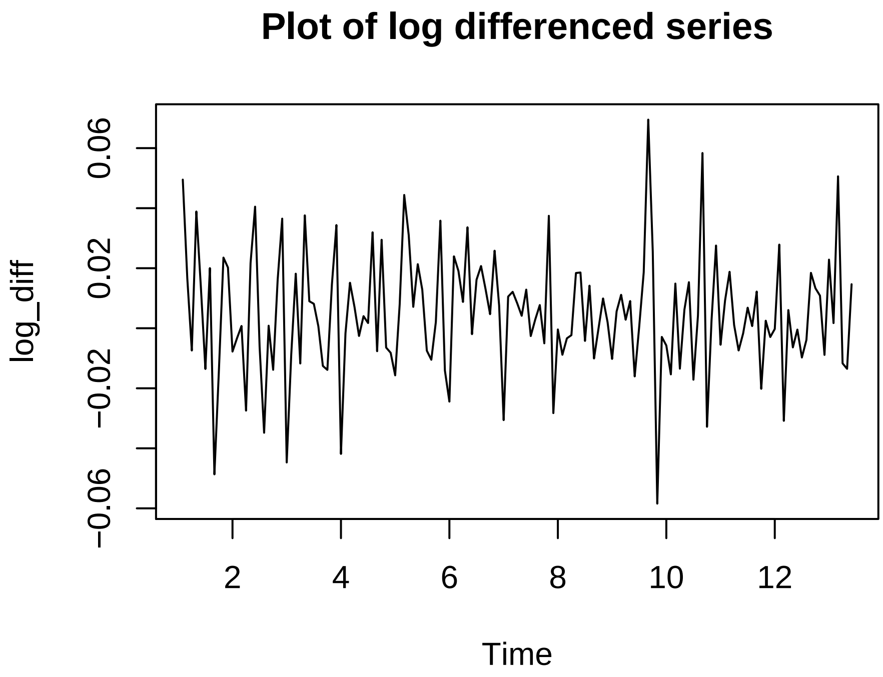
```
lambda #lambda value
```

```
## [1] 0.6767677
```

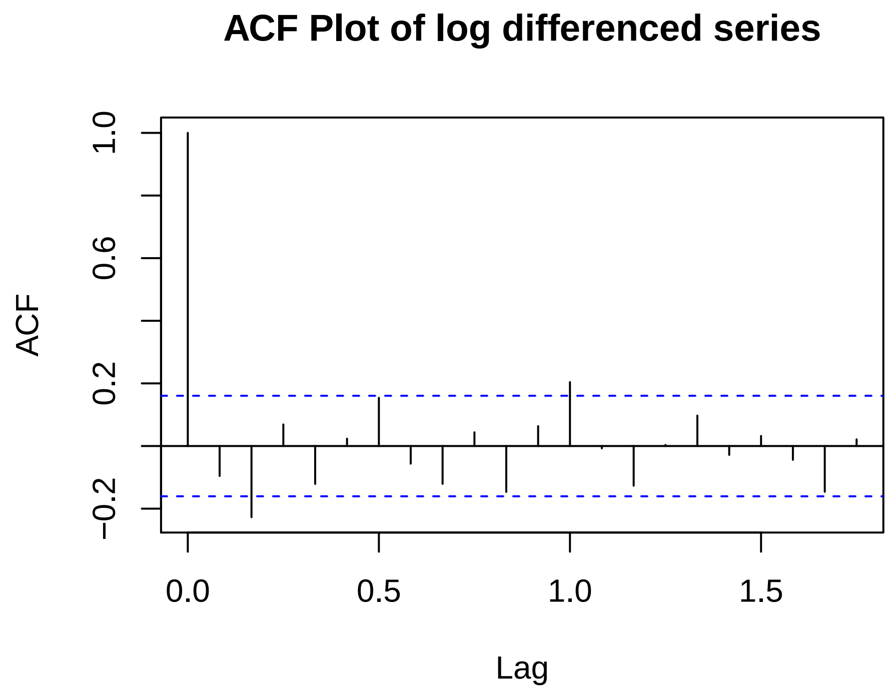
λ value	Appropriate Transformation
$\lambda = 1$	No substantive transformation
$\lambda = 0.5$	Square root plus linear transformation
$\lambda = 0$	Natural Log
$\lambda = -1$	Inverse plus 1

From the λ value we obtained above, the Box-Cox transformation method suggests that no further transformations are needed. However, based on the random walk and integrated behavior of the ACF of our `original series`, I think it is also sensible to perform `log`-differencing to the original time series (rather than just doing a classical decomposition). Applying the natural `log` transformation will also help make the variance of the original series to be more consistent as the original series has an exponential or geometric-like growth rate. First order differencing would also account for the potential seasonality presented in our series (which is very likely to be yearly difference). I will then compare the AIC/BIC again for the log-differenced series the MA(1) model of the remainder series with the same `astsa::sarima()` function to ensure consistency across results. The plots generated for the log differencing would be as follows:

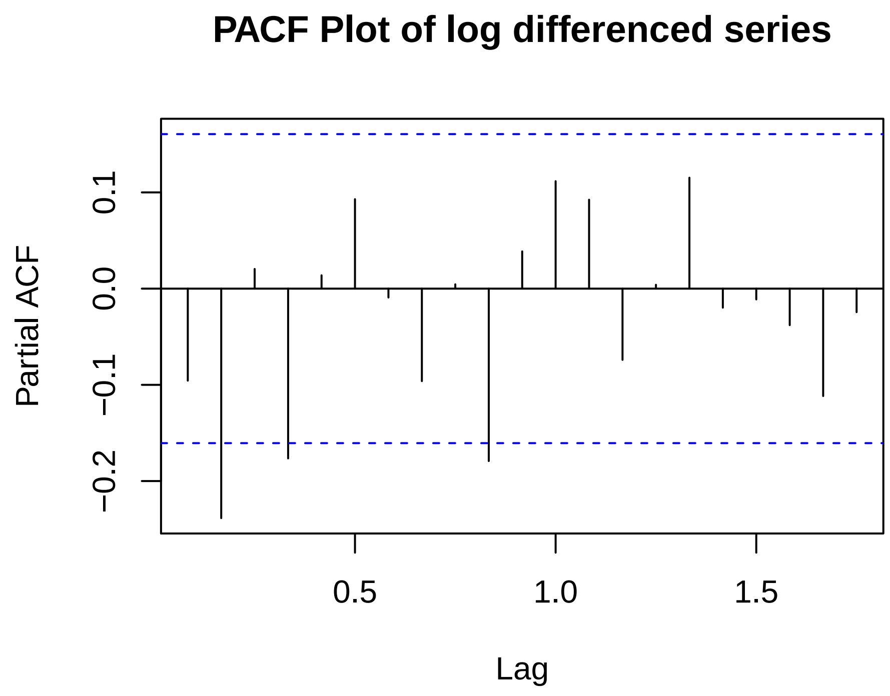
```r
log_diff = diff(log(ts))
plot(log_diff, main="Plot of log differenced series")
```



```r
acf(log_diff, main="ACF Plot of log differenced series")
```

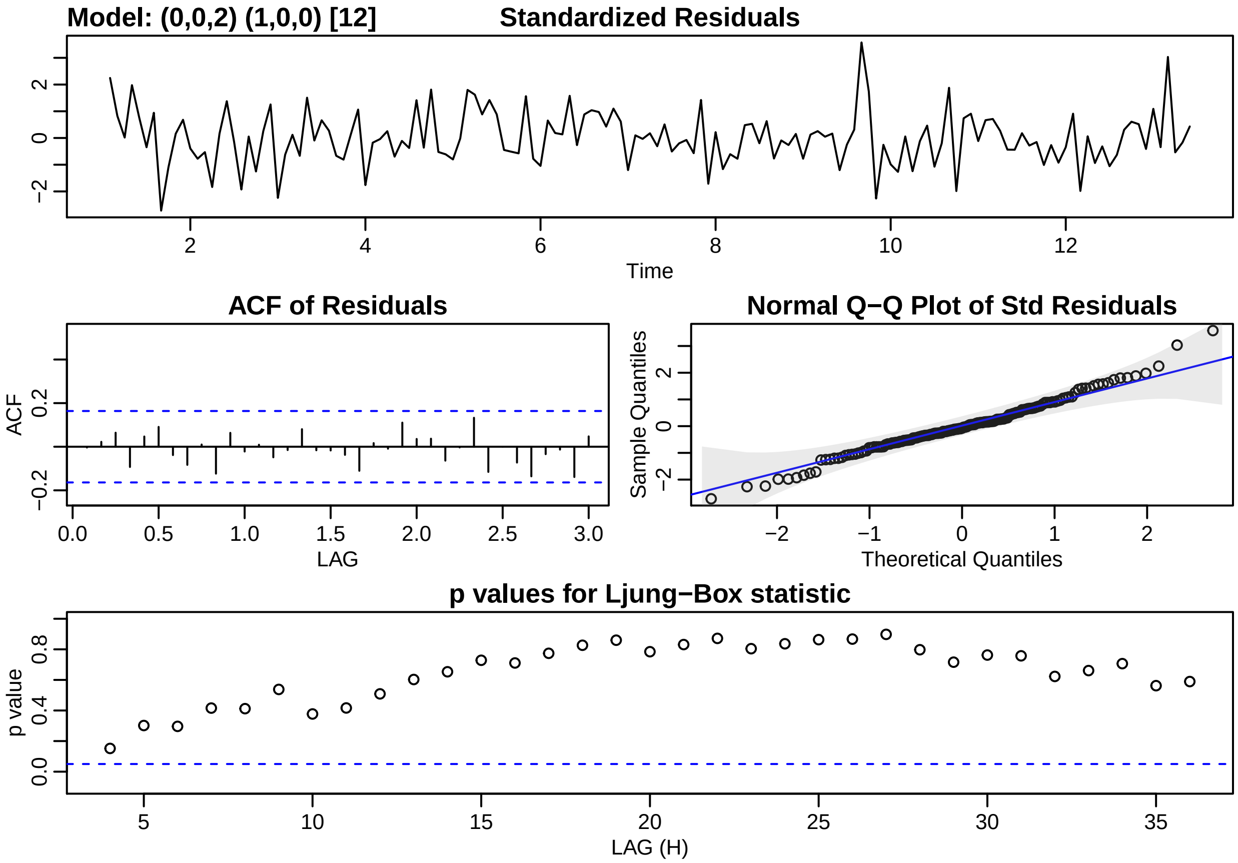


```r
pacf(log_diff, main="PACF Plot of log differenced series")
```



First, we could see that the log-differenced plot behaves like a stationary model with a constant mean and variance with slight indication of seasonality (which will be handled in a stochastic fashion in the SARIMA model we will fit later). I believe the ACF cuts off at lag 2, although there are some indication of correlation at lag 12 (but having a SARIMA model with order 12 would make our model overly complex and subjected to overfitting). The PACF also more or less tails off, making this an MA(2) model with first-order difference or a SARIMA$(0, 0, 2) \times (1, 0, 0)_{12}$ model. Now, we will perform some diagonistic checking as follows:

```r
sarima(log_diff,0,0,2,1,0,0,12)
```


Model: (0,0,2) (1,0,0) [12]

Standardized Residuals

ACF of Residuals

Normal Q−Q Plot of Std Residuals

p values for Ljung−Box statistic

```r
sarima_model <- Arima(log_diff, order=c(0,0,2), seasonal=c(1,0,0))
Box.test(sarima_model$resid, lag = 24, type = c("Ljung-Box"), fitdf = 8)$p.value
```

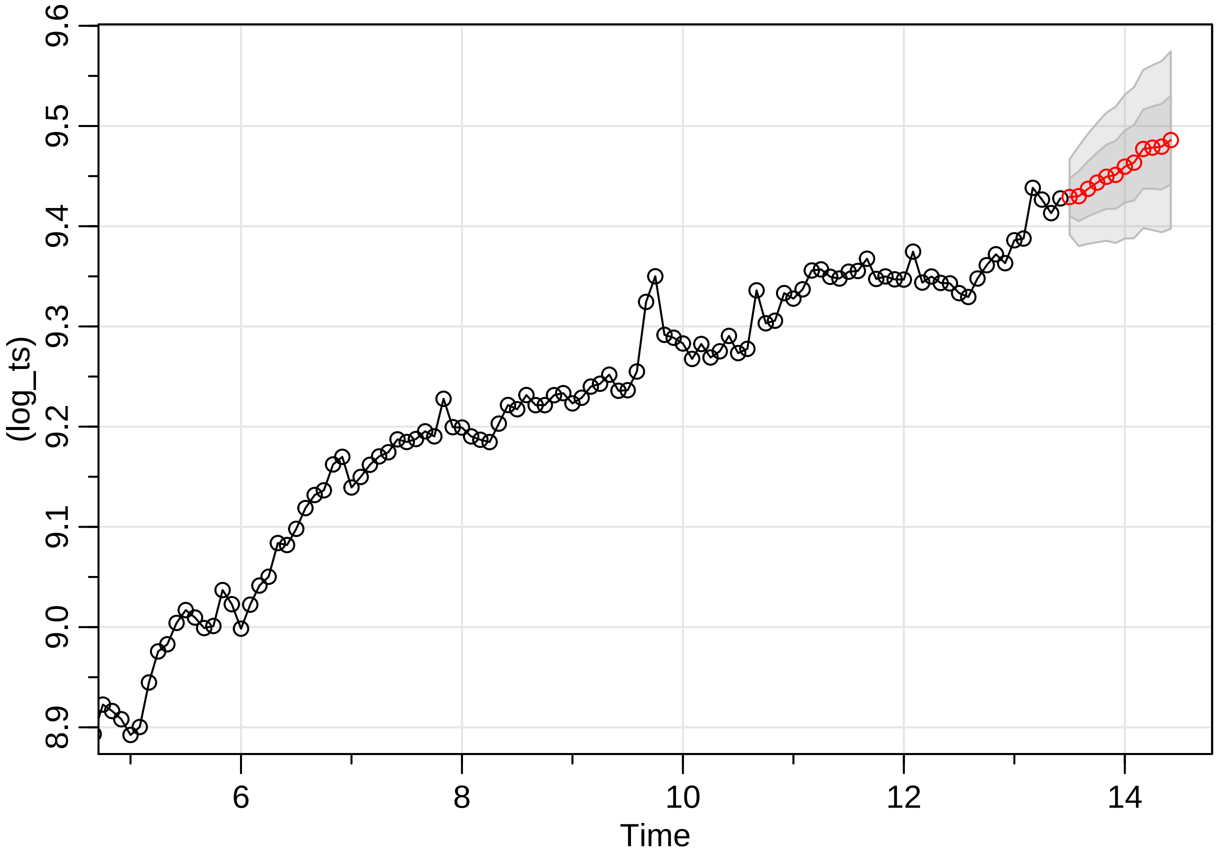
We can see from the ACF plot of the residuals are all within the 95% confidence intervals, indicating that there is no correlation between the residuals (suggesting a good fit of the model). The Normal Q-Q plot also behaves a lot better than the previous A2(2) or MA(1) model suggested from the classical decomposition, as there are no more visible outliers. The p-values are also all above the 5% confidence level (blue dashed line) as well, indicating that the model has no serial correlation with a 95% confidence level. This is also reinforced by the Ljung-Box model adequacy test as we have a p-value of 0.5447, to conlcude that the residuals are not serially correlated at the 95% confidence level. Finally, the AIC value is -5.03, AICc is -5.02, and BIC is -4.92998, which is also all significantly better than the previous MA(1) model. Hence, this model behaves the best from all 3 models in terms of diagnositic plots and AIC/BIC values. So, I will attempt to perform the 12-step-ahead forecast using this model.

Forecast (in `log` scale) using `sarima.for()` :

```r
#used log(ts) (without differencing) to ensure prediction is in the same scale
log_ts = log(ts)

#manually changed d=1 to account for difference
```

```
log_forecasts = sarima.for((log_ts),0,1,2,1,0,0,12, n.ahead=12)
```



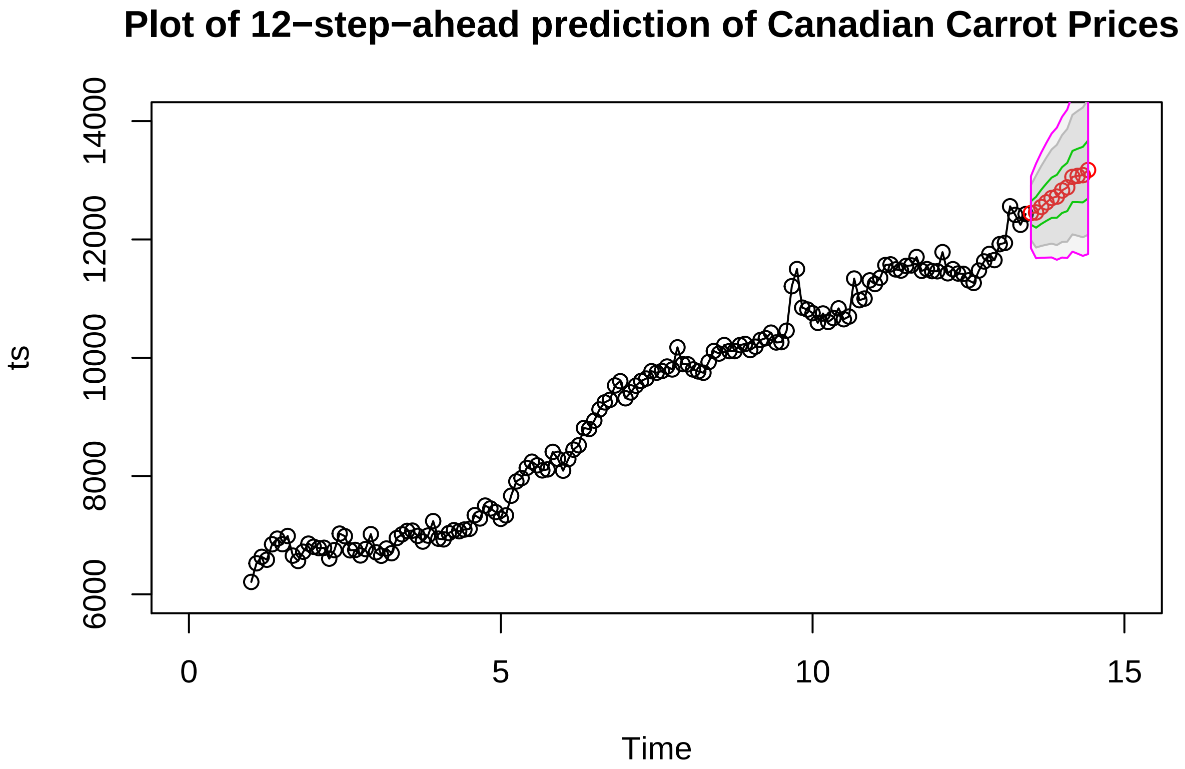
Original Scale Forecast plot (for 1 Prediction Error (captures 68% variability), 2 P.E. (95%), and 3 P.E. (99%)):

```
# Transform back to original scale forecasts
#2 P.E.
log_upper = log_forecasts$pred + 1.96 * log_forecasts$se
log_lower = log_forecasts$pred - 1.96 * log_forecasts$se
U = exp( log_upper )
L = exp( log_lower )

#1 P.E.
log_upper2 = log_forecasts$pred + 0.84 * log_forecasts$se
log_lower2 = log_forecasts$pred - 0.84 * log_forecasts$se
U2 = exp( log_upper2 )
L2 = exp( log_lower2 )

#3 P.E.
log_upper3 = log_forecasts$pred + 2.58 * log_forecasts$se
log_lower3 = log_forecasts$pred - 2.58 * log_forecasts$se
U3 = exp( log_upper3 )
L3 = exp( log_lower3 )
```

```
plot(ts, xlim=c(0,15), ylim=c(6000, 14000), type = 'o',
     main = "Plot of 12-step-ahead prediction of Canadian Carrot Prices ")
lines(exp( log_forecasts$pred), col = 2, type = 'o' )
    xx = c(time(U), rev(time(U)))
    yy = c(L, rev(U))
    xx2 = c(time(U2), rev(time(U2)))
    yy2 = c(L2, rev(U2))
    xx3 = c(time(U3), rev(time(U3)))
    yy3 = c(L3, rev(U3))
    polygon(xx, yy, border = 8, col = gray(0.6, alpha = 0.2)) #gray
    polygon(xx2, yy2, border = 3, col = gray(0.6, alpha = 0.1)) #green
    polygon(xx3, yy3, border = 6, col = gray(0.6, alpha = 0.1)) #purple
```



However, although this $\text{SARIMA}(0,0,2) \times (1,0,0)_{12}$ is the best model out of the 3 (compared to the classical decomposition models), there are still some limitations for this model. For example, when we choose the “simpler order” from the PACF and ACF of the log-differenced time series plots, we are subjected to bias. In other words, although there are advantages of choosing a simpler model with lower order of p and q (i.e., to avoid overfitting), we are also losing some information as one could always argue that there are indeed some sort of autocorrelation at the higher lags (since the PACF and ACF are indeed slightly above the 95% intervals at this higher lags). This might also be due to the lack of data we have since we only have 150 observations. If we were to obtain more data, the ACF and PACF plots might behave better and more consistently since having more data will give us stronger evidence of whether or not there is indeed a strong seasonal effect or indeed some significant correlation at higher lags, which will lead to a better time series model.