London Stock Price Report

Shi Fan Jin

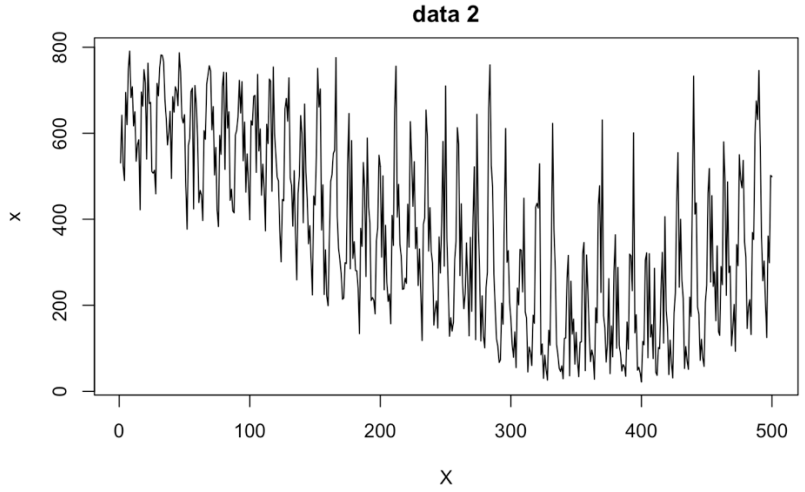
Dataset: data # 2

Introduction

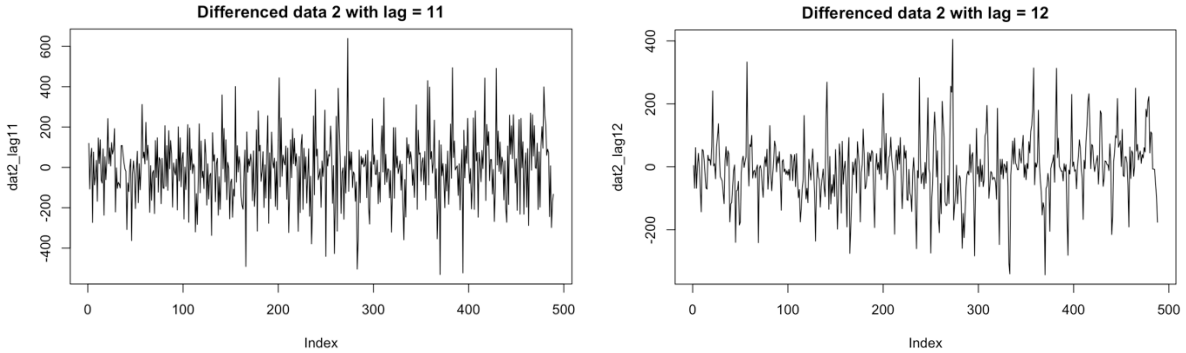
The report is based on the second dataset. The main purpose of this report is to find the 10 future predictions based on a fitted model. The outline of the report will be divided into three parts: Exploratory data analysis, Identifying suitable ARIMA model, Model fitting and forecasting.

A. Exploratory data analysis

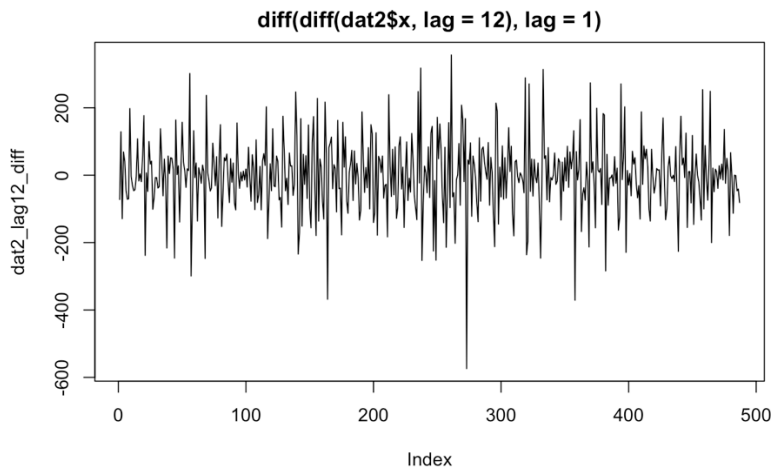
First, let's take a look at the plot of the original dataset.



The plot shows clear non-stationarity since the mean is not around zero, and the variation is not stable. It is obvious that there are some strong seasonal components, and we should try to remove them using differencing. From the plot, notice that there are about 8 or 9 peeks in each 100 period of x. This indicates that each cycle of the peek is about the length 11 or 12, which we can get by calculating 100/9 = 11.11 and 100/8 = 12.5 (rounded down). In this case, we can try to difference the data with either lag = 11 or lag = 12.



Looking at the plots, the plot with lag = 12 still seems to have a seasonal factor, so we should try to difference it again. Try differencing with lag = 1 this time.

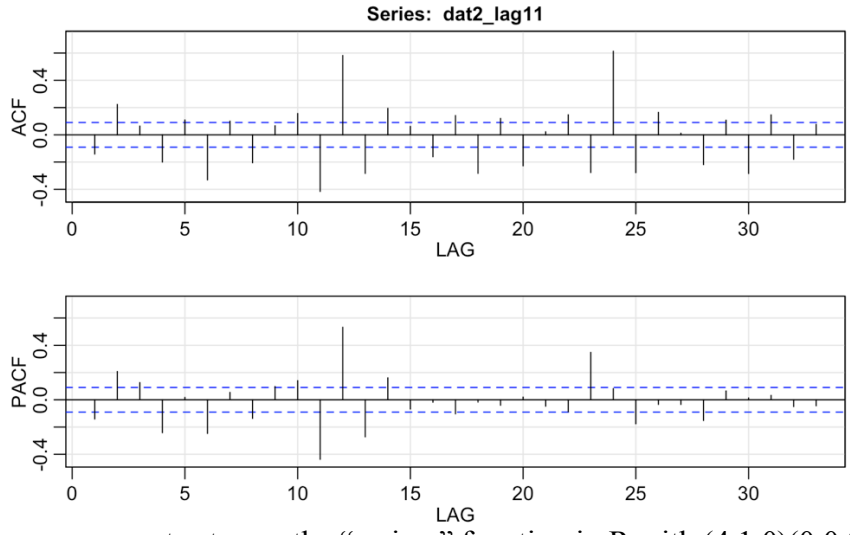


After differencing again with lag = 1, it looks more like a white noise and now everything seems stationary with the zero mean and stabilized variance.
Now, we have two transformed data as candidates for finding the fitted model:

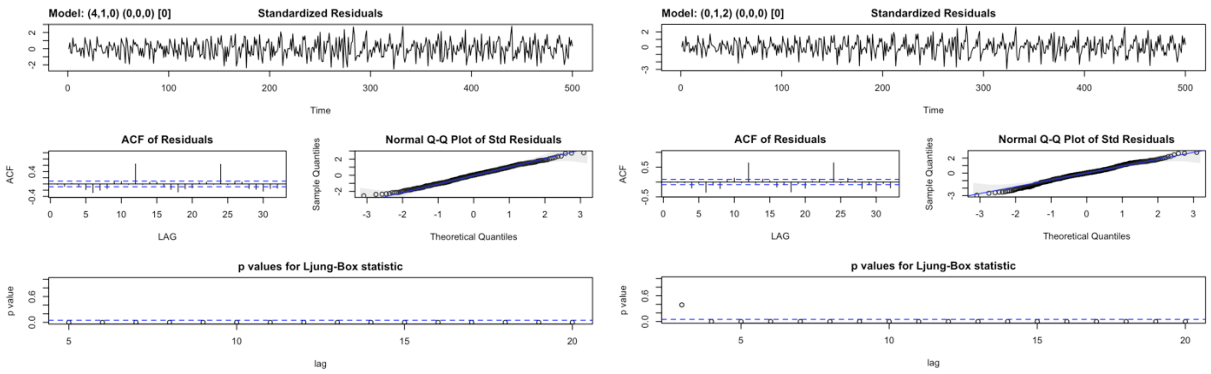
1. Differenced data with lag = 11 (Let's call this "dat2_lag11")
2. The first differenced of the seasonal (12) differenced data
   (Let's call this "dat2_lag12_diff")

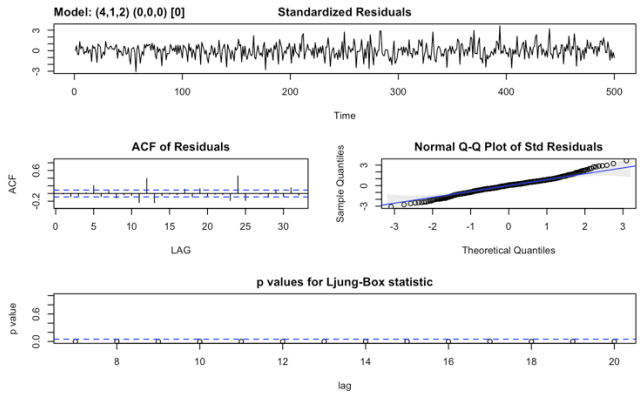
B. Identifying suitable ARIMA model

1. **Based on transformed data: "dat2_lag11"**
   Let's look at the ACF and PACF plots and see if we can find any useful clues. We can see that the ACF cuts off after lag 2, and the PACF cuts off after lag 4. This indicates the potential combination of (p, q) of the ARIMA process might be (4, 0), (0, 2), (4, 2).



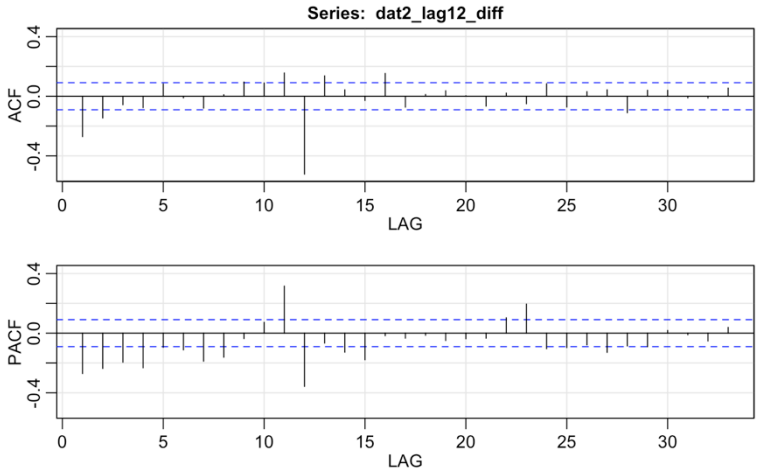
In this case, we can try to run the "sarima" function in R with (4,1,0)(0,0,0)[0], (0,1,2)(0,0,0)[0], (4,1,2)(0,0,0)[0] *(The order here is (p, d, q)(P, D, Q)[S])*



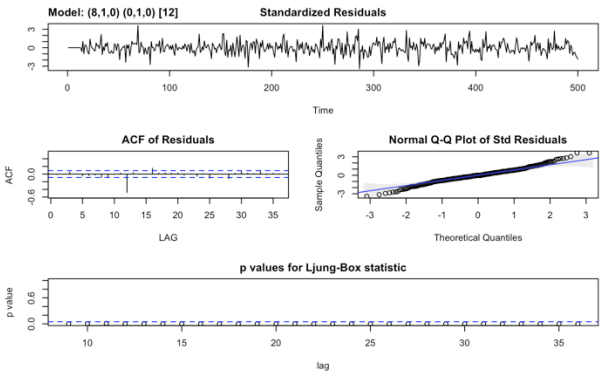


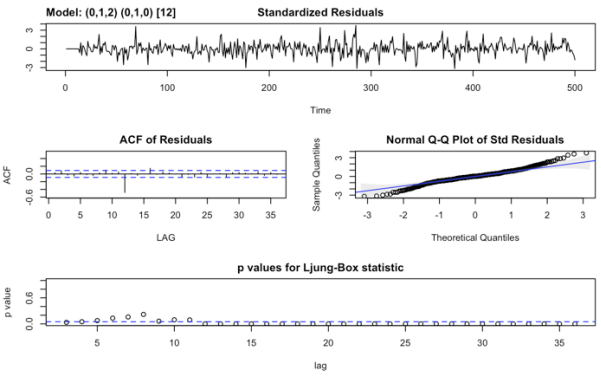
We can see right from the p-values of Ljung-Box statistic that all 3 models are not a good fit. Now let's switch to the other candidate and see if we can find a better model base on it.

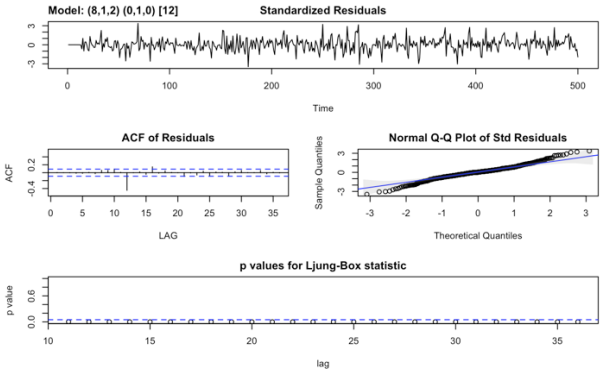
2. **Based on transformed data: "dat2_lag12_diff"**
   From the ACF and PACF plots, we can see that the ACF cuts off after lag 2, and the PACF cuts off after lag 8. This indicates the potential combination of (p, q) of the ARIMA process might be (8, 0), (0, 2), (8, 2).



In this case, we can try to run the "sarima" function in R with (8,1,0)(0,1,0)[12], (0,1,2)(0,1,0)[12], (8,1,2)(0,1,0)[12] *(The order here is (p, d, q)(P, D, Q)[S])*

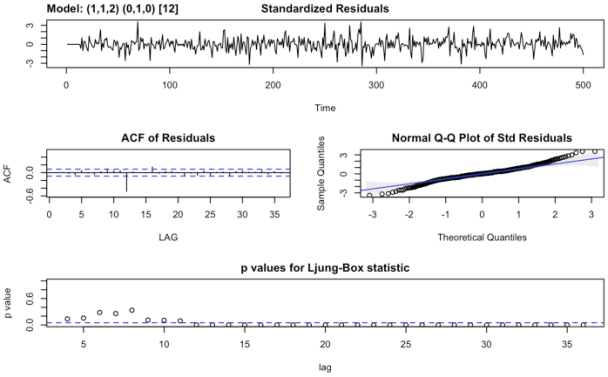






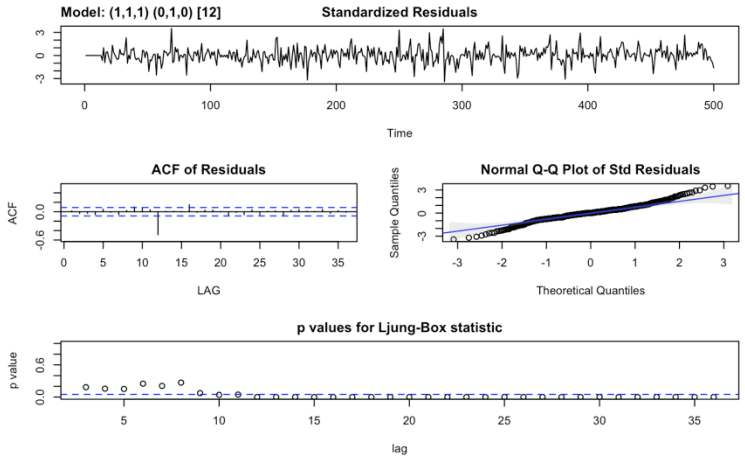
	Estimate	SE	t.value	p.value
ar1	-0.5704	0.0460	-12.4025	0.0000
ar2	0.3355	0.0536	6.2573	0.0000
ar3	-0.0646	0.0552	-1.1709	0.2422
ar4	-0.0733	0.0554	-1.3228	0.1865
ar5	0.0273	0.0553	0.4929	0.6223
ar6	0.0317	0.0551	0.5753	0.5654
ar7	-0.0482	0.0534	-0.9028	0.3671
ar8	-0.0528	0.0461	-1.1451	0.2527
ma1	0.0127	0.0120	1.0551	0.2919
ma2	-0.9873	0.0119	-82.8717	0.0000

Unfortunately, the Ljung-Box statistic shows that these are not good models. However, let's try calling "model$ttable" on the last model and see if we can get any clue. The output is on the bottom right corner of the above pictures. Looking at the p-values, it suggested us to try AR(1) or AR(2) instead of AR(8) since these p-values are less than 0.05. In this case, let's try (1,1,2)(0,1,0)[12].

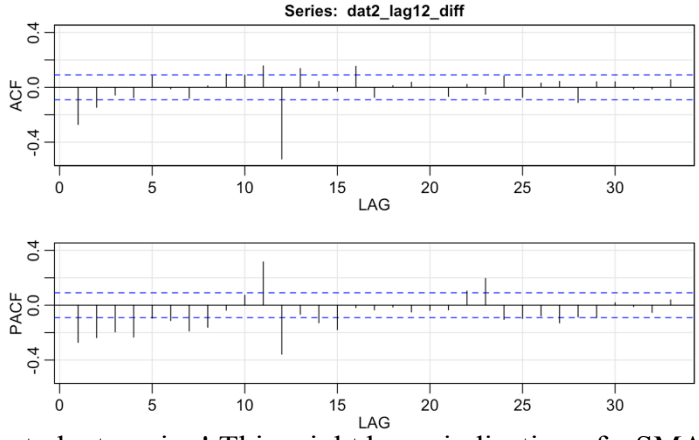


From the Ljung-Box, we can see that the p-values are starting to improve, this might be saying that we are going at the right direction. Let's try calling "model$ttable" again. Looking at the p-values, it suggested us to try MA(1) instead of MA(2), which is the (1,1,1)(0,1,0)[12] model.

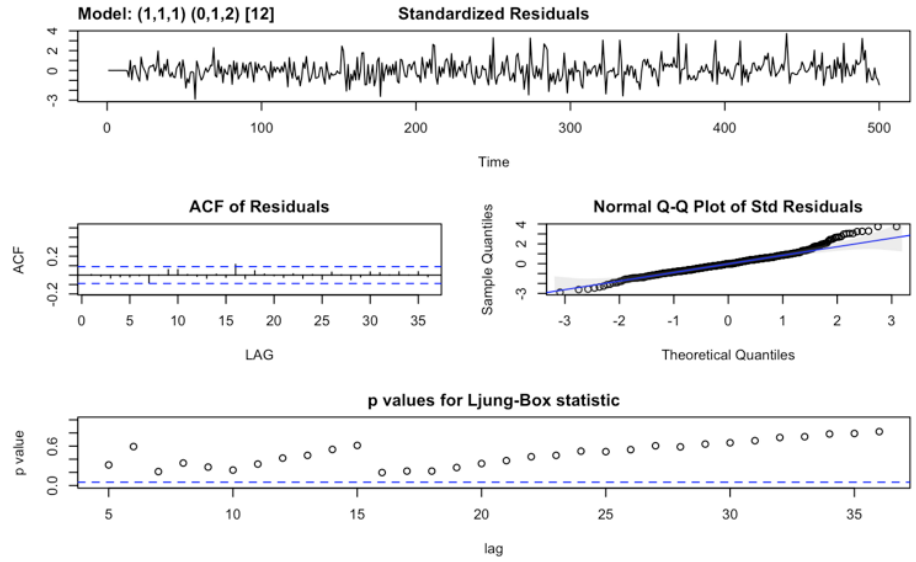
	Estimate	SE	t.value	p.value
ar1	0.2627	0.1121	2.3435	0.0195
ma1	-0.8556	0.1141	-7.4993	0.0000
ma2	-0.1327	0.1117	-1.1875	0.2356



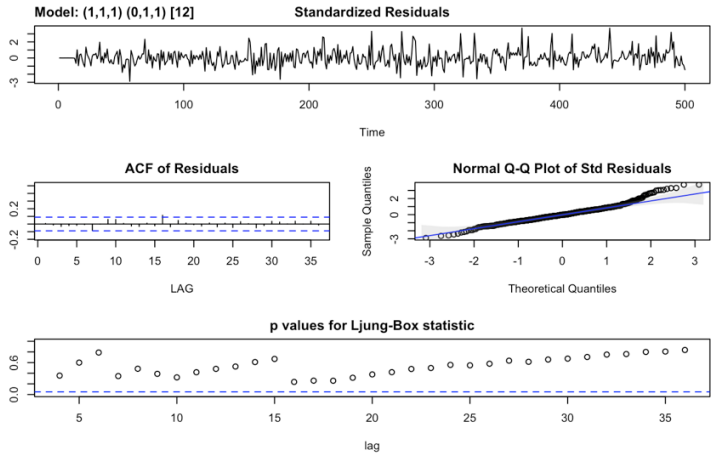
The output doesn't show any improvements. Let's go back to the ACF/PACF plot and check if we miss anything.



The PACF plot seems to be tapering! This might be an indication of a SMA(2) component! (ACF with sharp cutoff at lag 2, whereas the PACF is tapering) In this case, let's try to add this factor into the previous model. This gives us the following model: (1,1,1)(0,1,2)[12]



The output looks very good! The plot of standardized residuals looks like an i.i.d with mean 0 and variance 1 sequence. In addition, the Ljung-Box looks very good with a big improvement! Let's call "model$ttable" and see if we need to make any adjustments. It turns out that SMA(1) is better than SMA(2), so let's try the model: (1,1,1)(0,1,1)[12]



	Estimate	SE	t.value	p.value
ar1	0.3655	0.0477	7.6605	0.0000
ma1	-0.9276	0.0179	-51.8011	0.0000
sma1	-0.8726	0.0469	-18.5969	0.0000
sma2	0.0166	0.0464	0.3571	0.7212

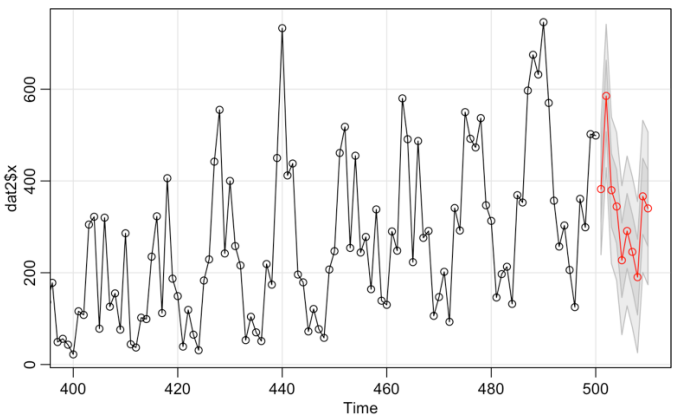
The result of (1,1,1)(0,1,1)[12] is also good! Now, the question will be which one is the better model to select.

To choose which model is better, let's compare the AIC, BIC, AICc scores. For model (1,1,1)(0,1,2)[12], the AIC, BIC, AICc is: 9.556031, 8.589747, 9.560273. As for the model (1,1,1)(0,1,1)[12], the AIC, BIC, AICc is: 9.552069, 8.577357, 9.556231. We can see that the second model has slightly better scores (smaller values). In this case, choose (1,1,1)(0,1,1)[12] as our final model for prediction.

C. Model fitting and forecasting

The coefficient is 0.3663 for the AR(1) component, -0.928 for MA(1), and -0.8598 for SMA(1). Now we can use the function "sarima.for" to predict the future 10 values.
In this case, the prediction of the 501 to 510 data is:

```
382.2626 585.2541 379.8301 344.3452 227.1662 291.0039 245.3819 190.3210 366.4625 340.1852
```




```
Coefficients:
         ar1      ma1     sma1
      0.3663  -0.9280  -0.8598
s.e.  0.0476   0.0178   0.0301

sigma^2 estimated as 5116:  log likelihood = -2779.67,  aic = 5567.34

$degrees_of_freedom
[1] 484

$ttable
     Estimate     SE  t.value p.value
ar1    0.3663 0.0476   7.6976       0
ma1   -0.9280 0.0178 -52.2281       0
sma1  -0.8598 0.0301 -28.5770       0

$AIC
[1] 9.552069

$AICc
[1] 9.556231

$BIC
[1] 8.577357
```

Appendix

#Data 2
```{r}
plot(dat2, type = "l", main = "data 2")
#Seems to have a seasonal factor and some kind of trend (quadratic, cubic)
```

##Data 2 Transformation
```{r}

# looks like there is a quadratic trend, take diff with differences = 2
dat2_diff2 = diff(dat2$x, differences = 2)
plot(dat2_diff2, type = "l", main = "Diff(dat2, differences = 2) to remove the quadratic trend")

plot(diff(diff(dat2$x)), type = "l", main = "diff(diff(dat2))")

##**diff(diff(dat2$x)) == diff(dat2$x, differences = 2)

# differenced with lag = 9
dat2_lag9 = diff(dat2$x, lag = 9)
plot(dat2_lag9, type = "l")
# there seems to be still a seasonal factor
dat2_lag9_10 = diff(dat2_lag9, lag = 10)
plot(dat2_lag9_10, type = "l")
# after differencing again with lag = 10


# looks like there's 9 peeks between 0 ~ 100, so each cycle would be 100/9,
# which is approximately 11.11, so choose lag = 11 for differencing for a try.
dat2_lag11 = diff(dat2$x, lag = 11)
plot(dat2_lag11, type = "l", main = "Differenced data 2 with lag = 11")
# This looks roughly stationary, with no seasonal component or linear trend.
# To recap, dat2_lag11 = (1 - B^11)x_t where x_t is our original data.

# looks like there's 8 peeks in each 100 period, cycle would be 100/8 = 12.5
# Try lag = 12
dat2_lag12 = diff(dat2$x, lag = 12)
plot(dat2_lag12, type = "l", main = "Differenced data 2 with lag = 12")
# This looks roughly stationary, with no seasonal component or linear trend.
# To recap, dat2_lag11 = (1 - B^12)x_t where x_t is our original data.
dat2_lag12_diff = diff(dat2_lag12)
plot(dat2_lag12_diff, type = "l", main = "diff(diff(dat2$x, lag = 12), lag = 1)")
# This looks roughly stationary, with no seasonal component or linear trend.
# To recap, dat2_lag11 = (1 - B^12)(1 - B)x_t where x_t is our original data.
```

CANDIDATES:
dat2_lag11:
  diff(dat2, lag = 11)
  # Differenced of lag(1)
dat2_lag12:
  diff(dat2, lag = 12)
  # Differenced of lag(12)
dat2_lag12_diff:
  diff(diff(dat2, lag = 12), lag = 1)
  # The first difference of the seasonal(12) differenced of data
```{r}
dat2_lag11 = diff(dat2$x, lag = 11)
plot(dat2_lag11, type = "l", main = "Differenced data 2 with lag = 11")

dat2_lag12 = diff(dat2$x, lag = 12)
plot(dat2_lag12, type = "l", main = "Differenced data 2 with lag = 12")

dat2_lag12_diff = diff(dat2_lag12)
plot(dat2_lag12_diff, type = "l", main = "diff(diff(dat2$x, lag = 12), lag = 1)")
```

###Check the candidates if stationary using ACF
```{r}
acf2(dat2_lag11)
# potential p = 4, q = 2
# potential combination of (p, q) = (4, 0), (0, 2), (4, 2)
# HOWEVER, the PACF seems to be tapering, so in this case maybe it indicates a SMA(2)
# NO SECOND DIFF WITH SEASONAL LAG >>> CANT USE SAR OR SMA!!!

acf2(dat2_lag12)
# potential p = 1, q = 2
# potential combination of (p, q) = (1, 0), (0, 2), (1, 2)
# HOWEVER, the ACF seems to be tapering, so in this case maybe it indicates a SAR(1)
# NO SECOND DIFF WITH SEASONAL LAG >>> CANT USE SAR OR SMA!!!

acf2(dat2_lag12_diff)
# potential p = 8, q = 2
# potential combination of (p, q) = (8, 0), (0, 2), (8, 2)
# THIS ONE'S PACF DOESN'T SEEM TO BE GOOD (DOESN'T HAVE OBVIOUS CUTOFF
IN THE FRONT)
# PACF SEEMS LIKE TAPPERING, MAYBE INCLUDE SMA(2)?
```

##Finding Fitted Model

```{r}
##################  FROM dat2_lag11
## potential combination of (p, q) = (4, 0), (0, 2), (4, 2)
#### (4,1,0)(0,0,0)[0]
model2A = sarima(dat2$x,p=4,d=1,q=0,P=0,D=0,Q=0,S=0)   #BAD
model2A$ttable
#### (0,1,2)(0,0,0)[0]
model2A_1 = sarima(dat2$x,p=0,d=1,q=2,P=0,D=0,Q=0,S=0)   #BAD
model2A_1$ttable
#### (4,1,2)(0,0,0)[0]
model2A_2 = sarima(dat2$x,p=4,d=1,q=2,P=0,D=0,Q=0,S=0)   #BAD
model2A_2$ttable



################## FROM dat2_lag12
## potential combination of (p, q) = (1, 0), (0, 2), (1, 2)
#### (1,1,0)(0,0,0)[0]
model2B = sarima(dat2$x,p=1,d=1,q=0,P=0,D=0,Q=0,S=0)   #BAD
model2B$ttable
#### (0,1,2)(0,0,0)[0]
model2B_1 = sarima(dat2$x,p=0,d=1,q=2,P=0,D=0,Q=0,S=0)   #IMPROVING
model2B_1$ttable
#### (1,1,2)(0,0,0)[0]
model2B_2 = sarima(dat2$x,p=1,d=1,q=2,P=0,D=0,Q=0,S=0)   #IMPROVING
model2B_2$ttable



##################  FROM dat2_lag12_diff
## potential combination of (p, q) = (8, 0), (0, 2), (8, 2)
#### (8,1,0)(0,1,0)[12]
model2C = sarima(dat2$x,p=8,d=1,q=0,P=0,D=1,Q=0,S=12)   #BAD
model2C$ttable
#### (0,1,2)(0,1,0)[12]
model2C_1 = sarima(dat2$x,p=0,d=1,q=2,P=0,D=1,Q=0,S=12)   #IMPROVING
model2C_1$ttable
#### (8,1,2)(0,1,0)[12]
model2C_2 = sarima(dat2$x,p=8,d=1,q=2,P=0,D=1,Q=0,S=12)   #BAD
model2C_2$ttable
# FROM THE p-values, suggested AR(1) or AR(2)? instead of AR(8)
#### (1,1,2)(0,1,0)[12]
model2C_3 = sarima(dat2$x,p=1,d=1,q=2,P=0,D=1,Q=0,S=12)   #IMPROVING
model2C_3$ttable
# FROM THE p-values, MAYBE MA(1) instead of MA(2)?
#### (1,1,1)(0,1,0)[12]
```

```
model2C_4 = sarima(dat2$x,p=1,d=1,q=1,P=0,D=1,Q=0,S=12)   #IMPROVING
model2C_4$ttable
model2C_4
# STILL NOT GOOD ENOUGH, MAYBE GO BACK TO ACF2 TO SEE IF THERE IS SEASONAL (P,D,Q)?
# PACF SEEMS LIKE TAPPERING >>> MAYBE INCLUDE SMA(2)?
#### (1,1,1)(0,1,2)[12]
model2C_5 = sarima(dat2$x,p=1,d=1,q=1,P=0,D=1,Q=2,S=12)   #VERY GOOD!
model2C_5$ttable
model2C_5$AIC
model2C_5$BIC
model2C_5$AICc

#### (1,1,1)(0,1,1)[12]
model2C_6 = sarima(dat2$x,p=1,d=1,q=1,P=0,D=1,Q=1,S=12)   #VERY GOOD!!!!
######### BEST FOR NOW!!!!!
model2C_6$ttable
model2C_6$AIC
model2C_6$BIC
model2C_6$AICc

#### (1,1,1)(1,1,2)[12]
model2B_5 = sarima(dat2$x,p=1,d=1,q=1,P=1,D=1,Q=2,S=12)   #VERY GOOD!!!
model2B_5$ttable
model2B_5$AIC
model2B_5$BIC
model2B_5$AICc

################## FROM XEINAI
#### (1,1,1)(1,1,1)[12]
model2XN = sarima(dat2$x,p=1,d=1,q=1,P=1,D=1,Q=1,S=12)
model2XN$ttable
model2XN$AIC
model2XN$BIC
model2XN$AICc

#########################################################CHOOSE (1,1,1)(0,1,1)[12]
model2C_5$fit
```

##MODEL SELECTION
```{r}
#### (1,1,1)(0,1,1)[12]
#model2C_6 = sarima(dat2$x,p=1,d=1,q=1,P=0,D=1,Q=1,S=12)   #VERY GOOD!!!!
######### BEST FOR NOW!!!!!
#model2C_6$ttable
```

```
model2C_6$AIC
model2C_6$BIC
model2C_6$AICc

#### (1,1,1)(1,1,2)[12]
#model2B_5 = sarima(dat2$x,p=1,d=1,q=1,P=1,D=1,Q=2,S=12)   #VERY GOOD!!!
#model2B_5$ttable
model2B_5$AIC
model2B_5$BIC
model2B_5$AICc

################## FROM XEINAI
#### (1,1,1)(1,1,1)[12]
#model2XN = sarima(dat2$x,p=1,d=1,q=1,P=1,D=1,Q=1,S=12)
#model2XN$ttable
model2XN$AIC
model2XN$BIC
model2XN$AICc


########################LOOKING AT THE AIC/BIC/AICc, MODEL: (1,1,1)(0,1,1)[12] is the best! (Having the smallest values)
model2C_5 = sarima(dat2$x,p=1,d=1,q=1,P=0,D=1,Q=1,S=12)
```


##FORECASTING
```{r}
model2C_5

# USING (1,1,1)(0,1,1)[12]
pred2 <- sarima.for(xdata = dat2$x, n.ahead = 10, p=1,d=1,q=1,P=0,D=1,Q=1,S=12)
pred2$pred
plot(dat2$x, type = "l")

```