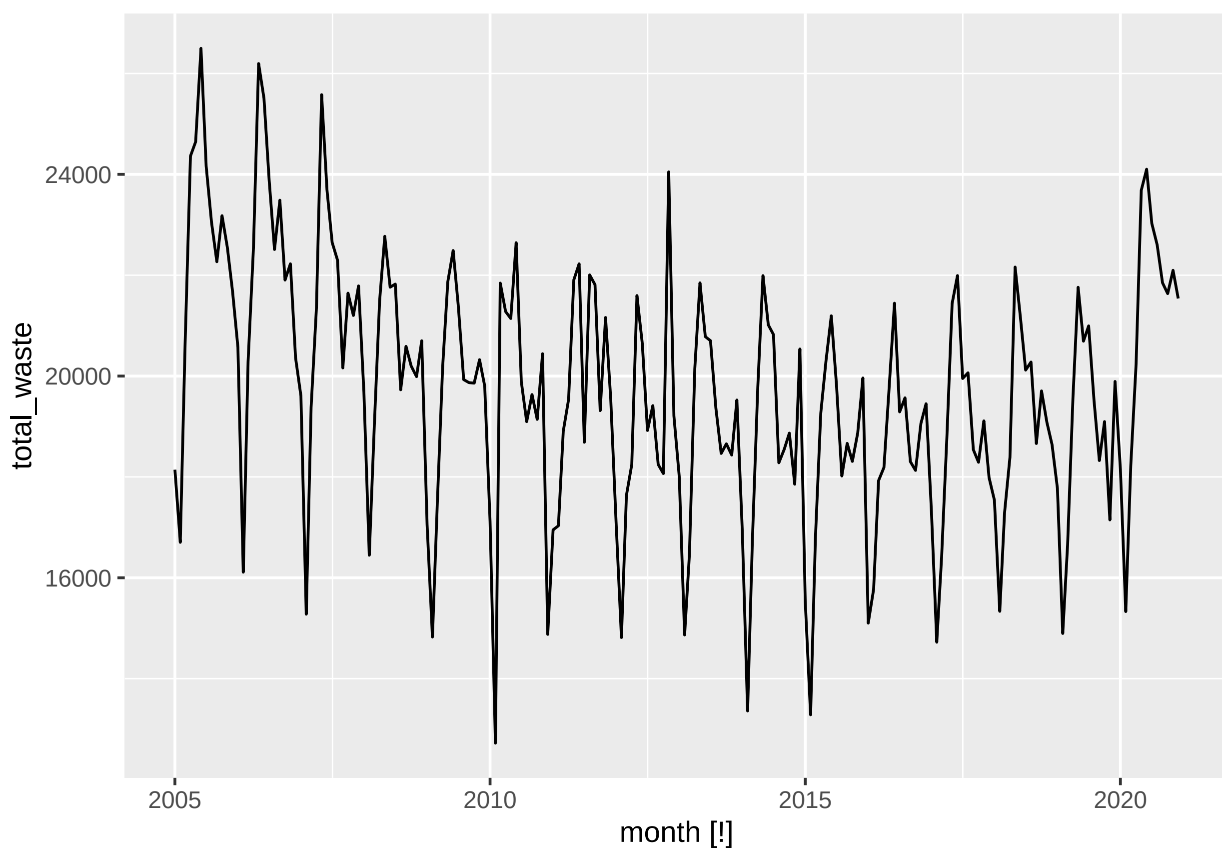
Staten Island Time Series

Daniel L.

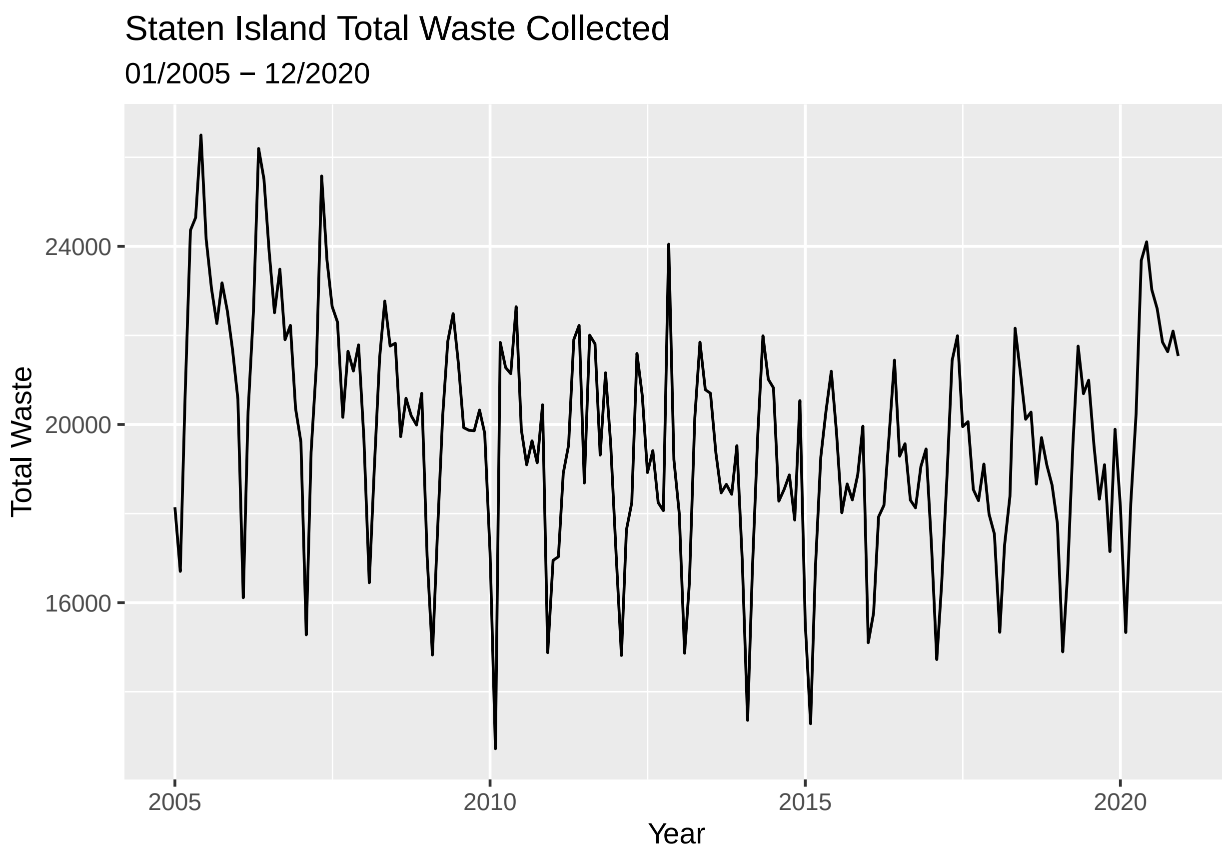
4/30/2022

```
autoplot(si_ts)
```

```
## Plot variable not specified, automatically selected `.vars = total_waste`
```




```
si_ts %>%
  ggplot(mapping = aes(x = month, y = total_waste)) + geom_line() +
  labs(x = "Year",
       y = "Total Waste",
       title = "Staten Island Total Waste Collected",
       subtitle = "01/2005 - 12/2020")
```



We see that the majority of the values are bounded b/w (15000, 26000)

KPSS Test for 'total_waste'
H_0 : The time series is trend stationary vs H_a : The time series is not trend stationary

If the p-value of the test is less than some significance level (e.g. $\alpha = .05$) then we reject the null hypothesis and conclude that the time series is not trend stationary.

```
si_ts %>% features(total_waste, unitroot_kpss)
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1     0.988        0.01
```

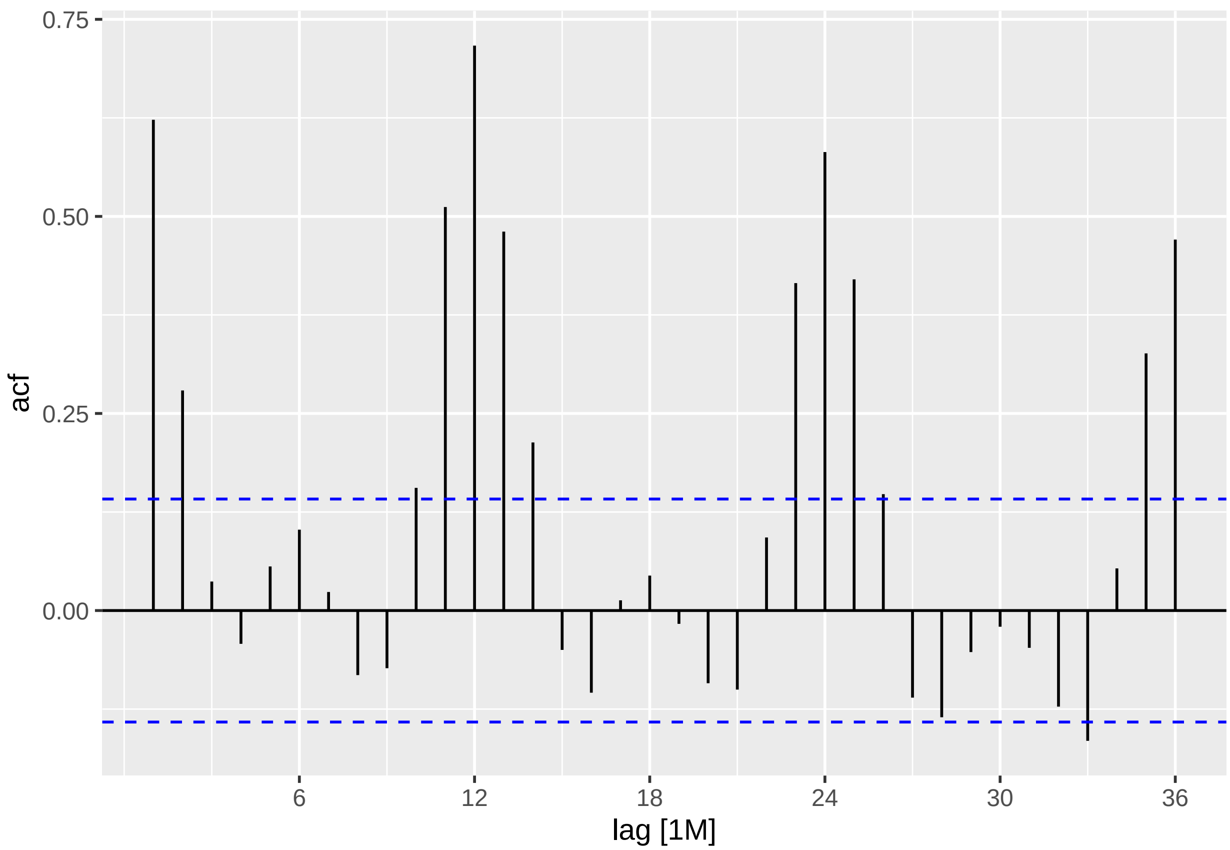
Using the KPSS test, we are returned a p-value of .01, We reject H_0

```
si_ts %>% features(diff1, unitroot_kpss)
```

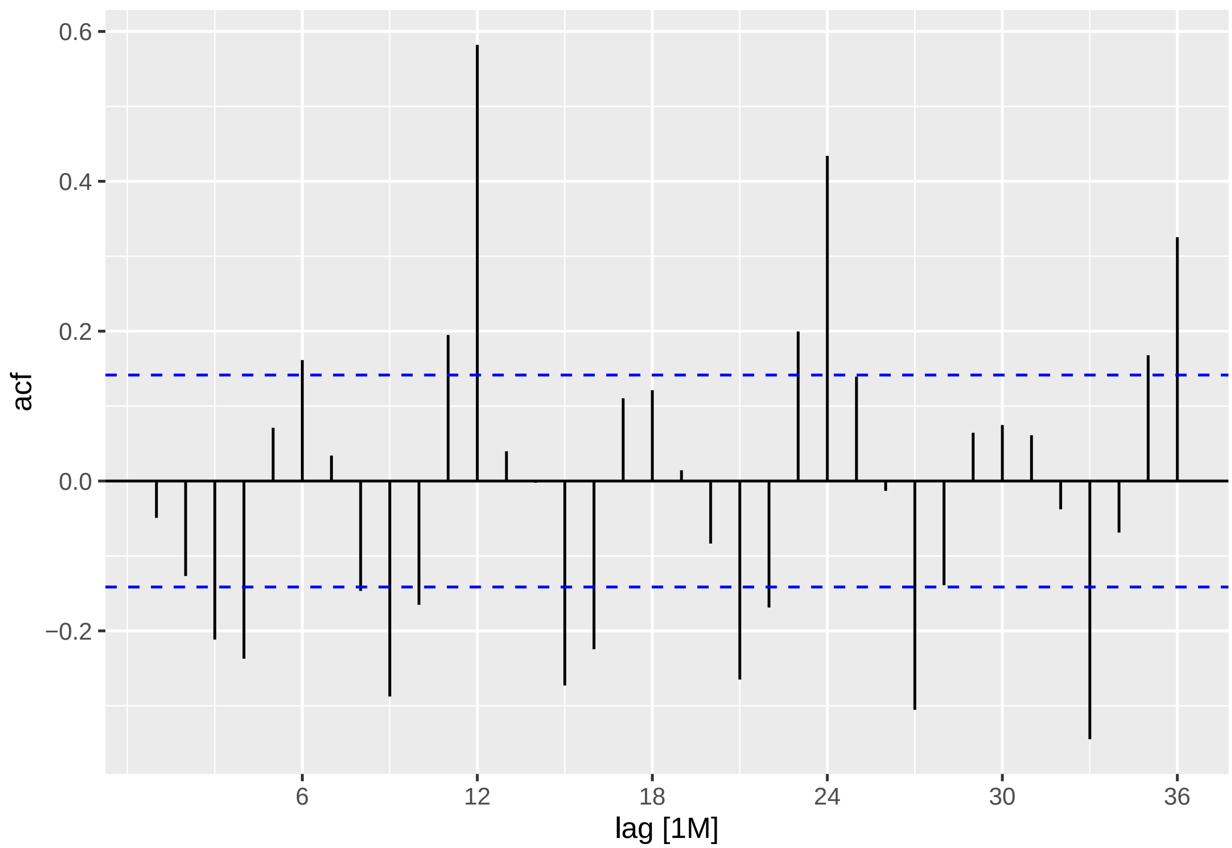
```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1    0.0192         0.1
```

Begin by looking at ACF and PACF of the total_waste and differenced values

```
si_ts2 %>%
  ACF(total_waste, lag_max = 36) %>%
  autoplot()
```

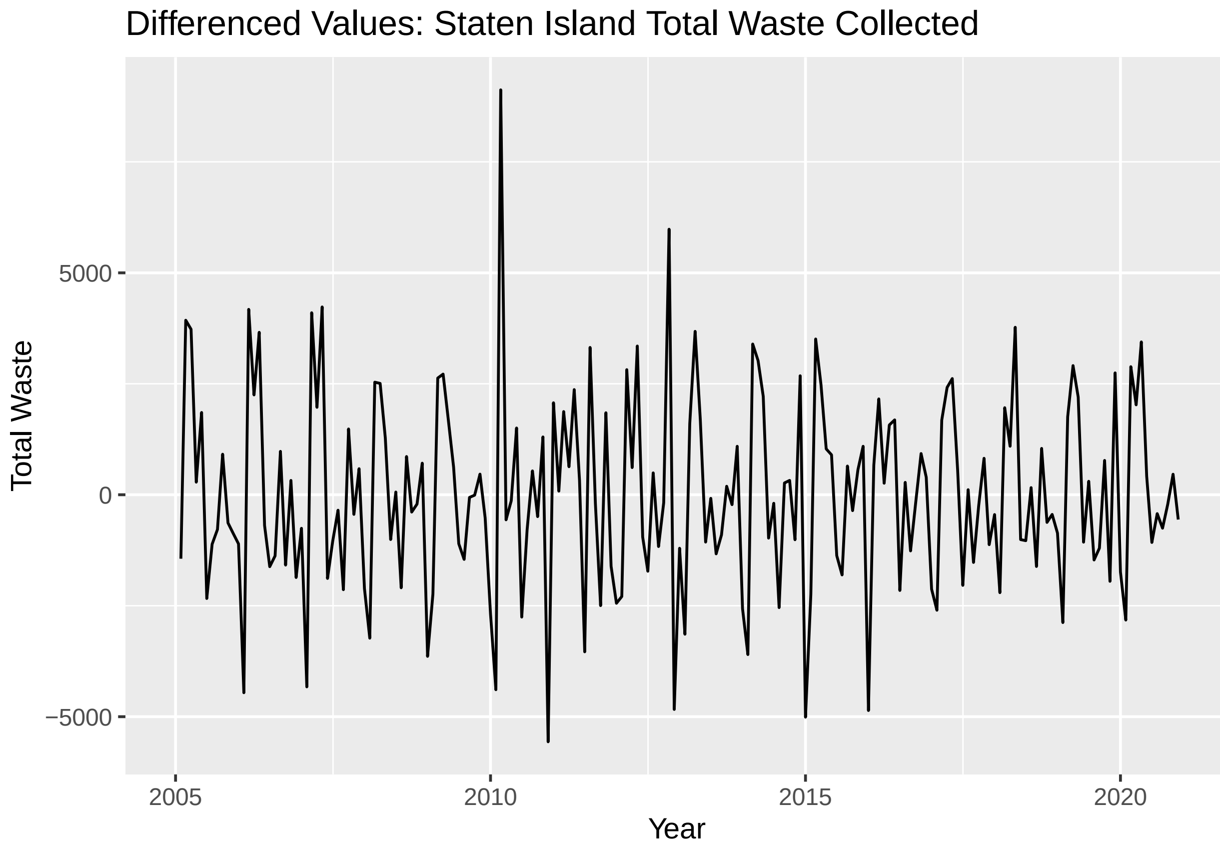



```
#acf of the differenced values
si_ts2 %>%
  ACF(diff1, lag_max = 36) %>%
  autoplot()
```



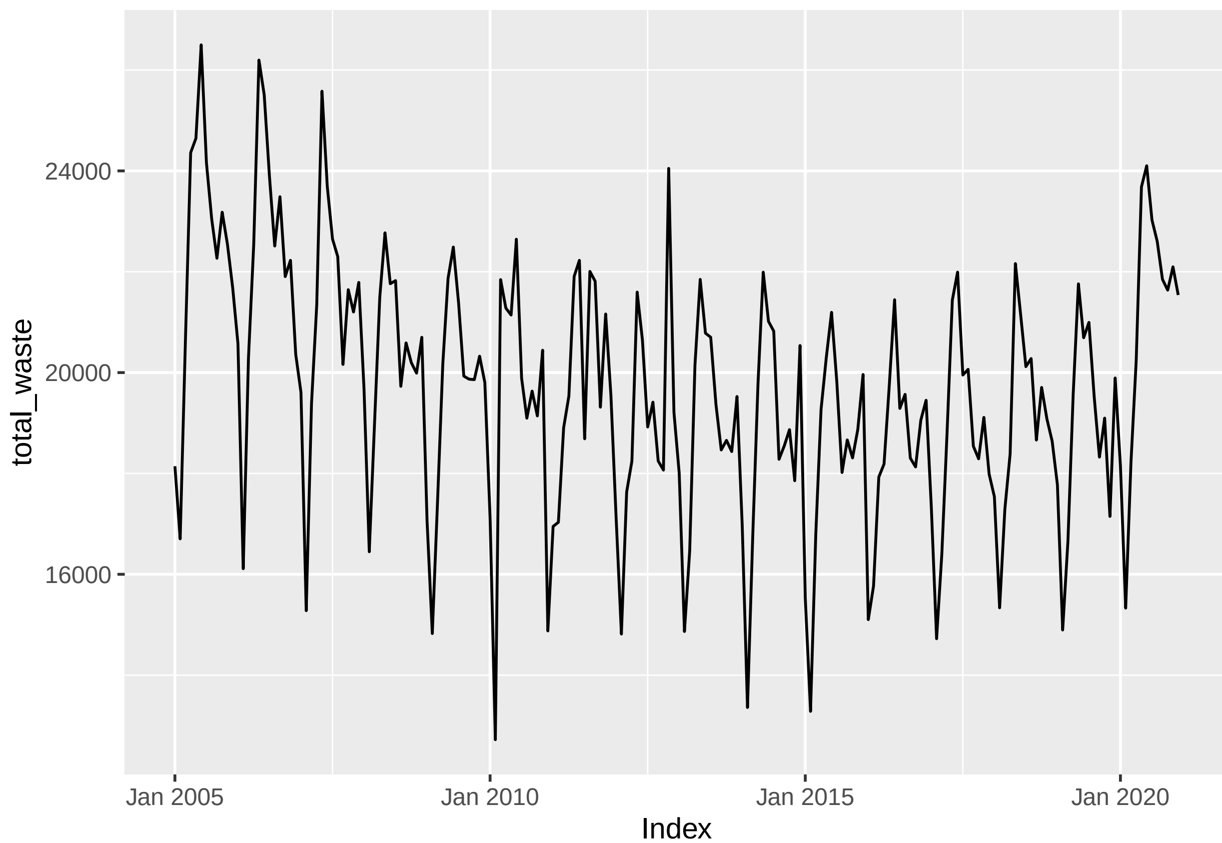

```
## Diff1 ggplot
si_ts3 %>%
  ggplot(mapping = aes(x = month, y = diff1)) + geom_line() +
  labs(x = "Year", y = "Total Waste", title = "Differenced Values: Staten Island Total Waste Collected")
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



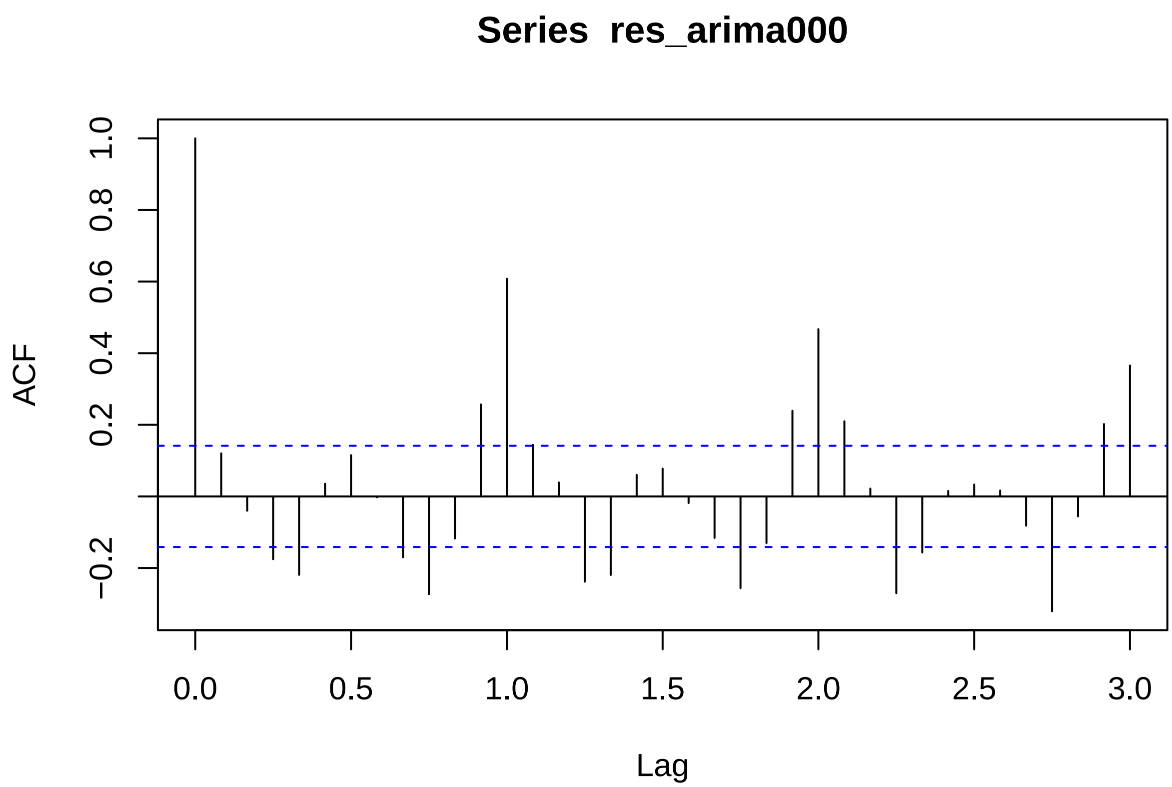
Creating models with zoo() and the arima package from stats()

```
# DSNY_third_staten_island[1]
DSNY_SI_zoo_ts <- ts(DSNY_third_staten_island[,2],
                     start = as.yearmon(DSNY_third_staten_island$month)[1],
                     frequency = 12)

autoplot(as.zoo(DSNY_SI_zoo_ts))
```

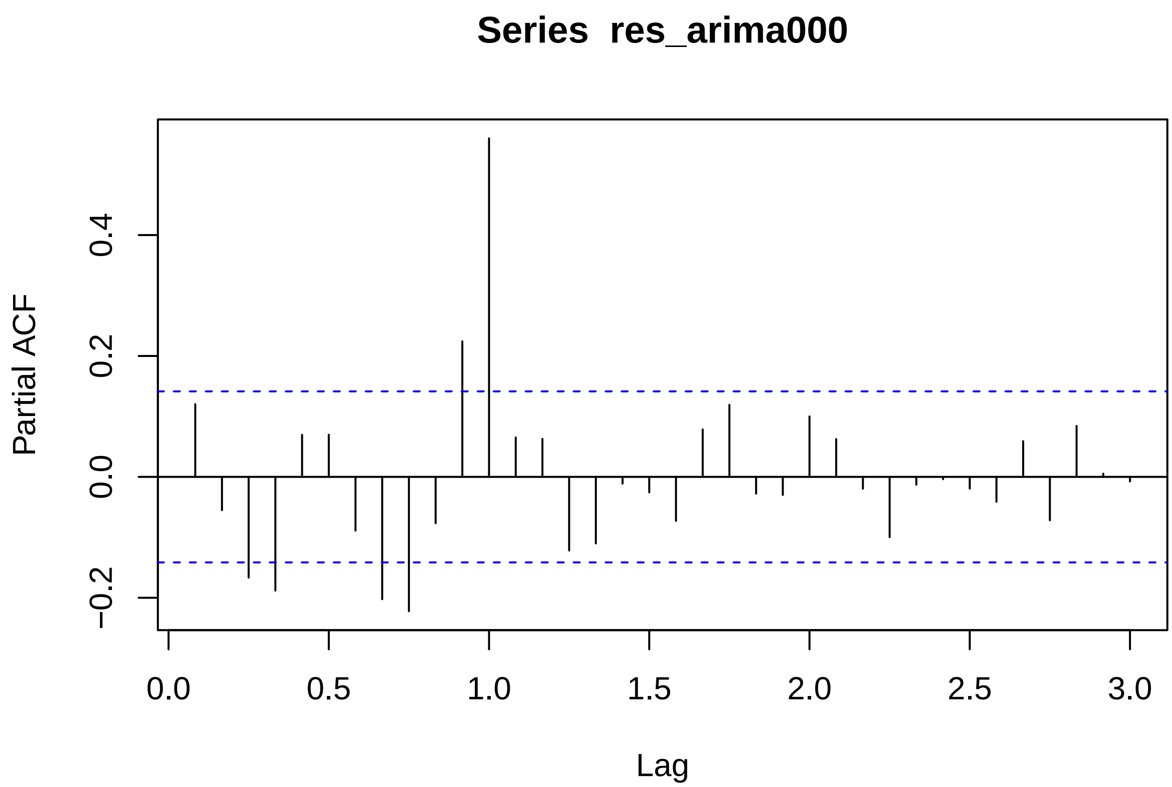


ARIMA(0,0,0) with constant

```
si_arima000_fit_cons <- si_ts2 %>%
  model(arima000_constant = ARIMA(total_waste ~ 1 +
                                    pdq(0,0,0)))

zoo_arima000_fit <- arima(DSNY_SI_zoo_ts, order = 1 + c(0,0,0))
#names(zoo_arima000_fit)
res_arima000 <- zoo_arima000_fit$residuals
acf(res_arima000, lag.max = 36)
```




```
pacf(res_arima000, lag.max = 36)
```




```
accuracy(si_arima000_fit_cons)[4]
```

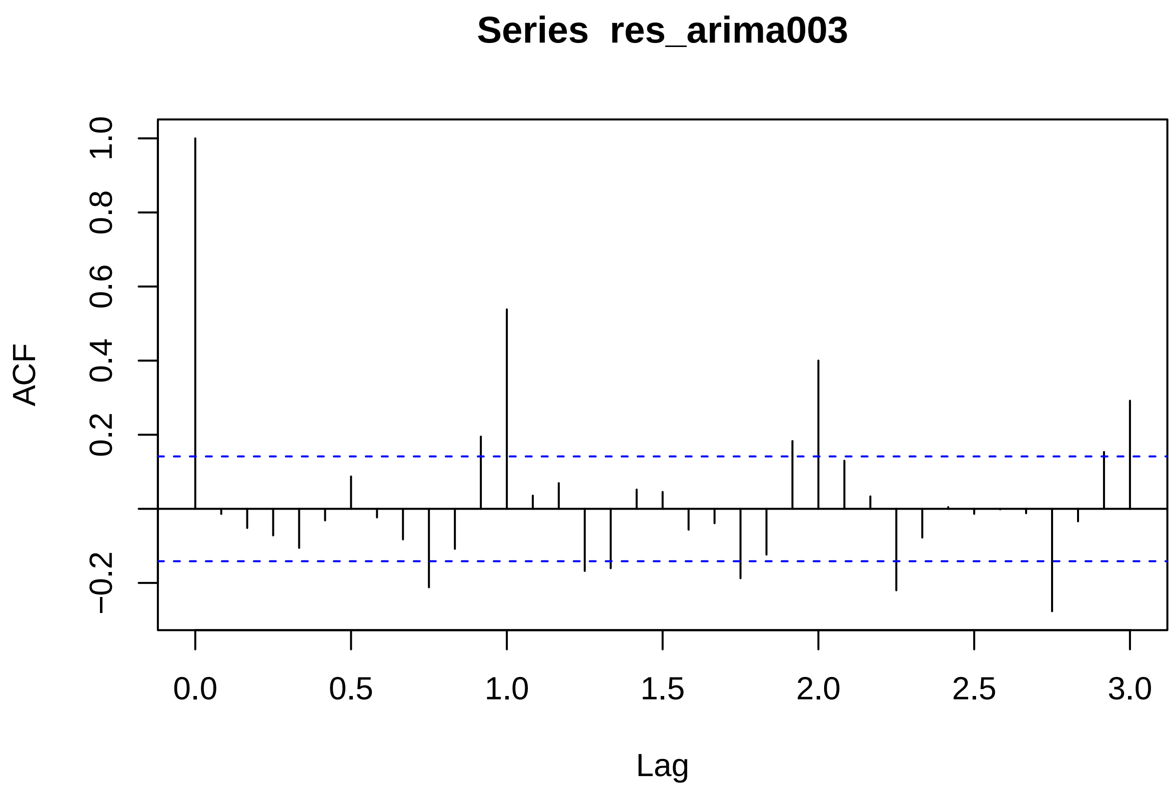
```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 1400.
```

From the first PACF plot, $ARIMA(0,0,0)$ we see significant values at lags 3 and 4. Lag 3 has a negative PACF value, indicating that we should begin with an MA(3) or MA(4) model on the undifferenced data. We also see that there is a significant positive value on lag 12. Because we know that our time series data is seasonal, we can look to add a seasonal AR parameter on the undifferenced data.

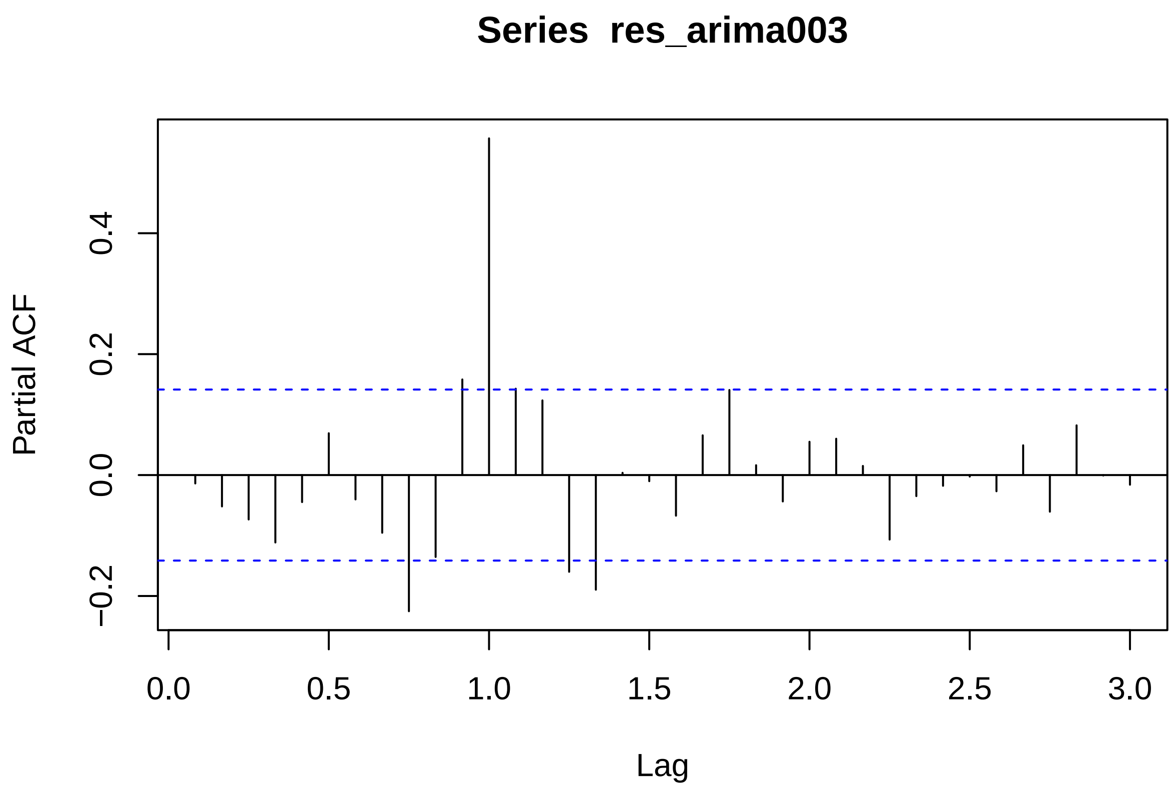
ARIMA(0,0,3) with constant

Let's begin with $ARIMA(0,0,3)$

```
si_arima003_fit_cons <- si_ts2 %>%
  model(arima002_constant = ARIMA(total_waste ~ 1 +
                                    pdq(0,0,3)))

zoo_arima003_fit <- arima(DSNY_SI_zoo_ts, order = 1 + c(0,0,3))
#names(zoo_arima000_fit)
res_arima003 <- zoo_arima003_fit$residuals
acf(res_arima003, lag.max = 36)
```




```
pacf(res_arima003, lag.max = 36)
```




```
accuracy(si_arima003_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 1251.
```

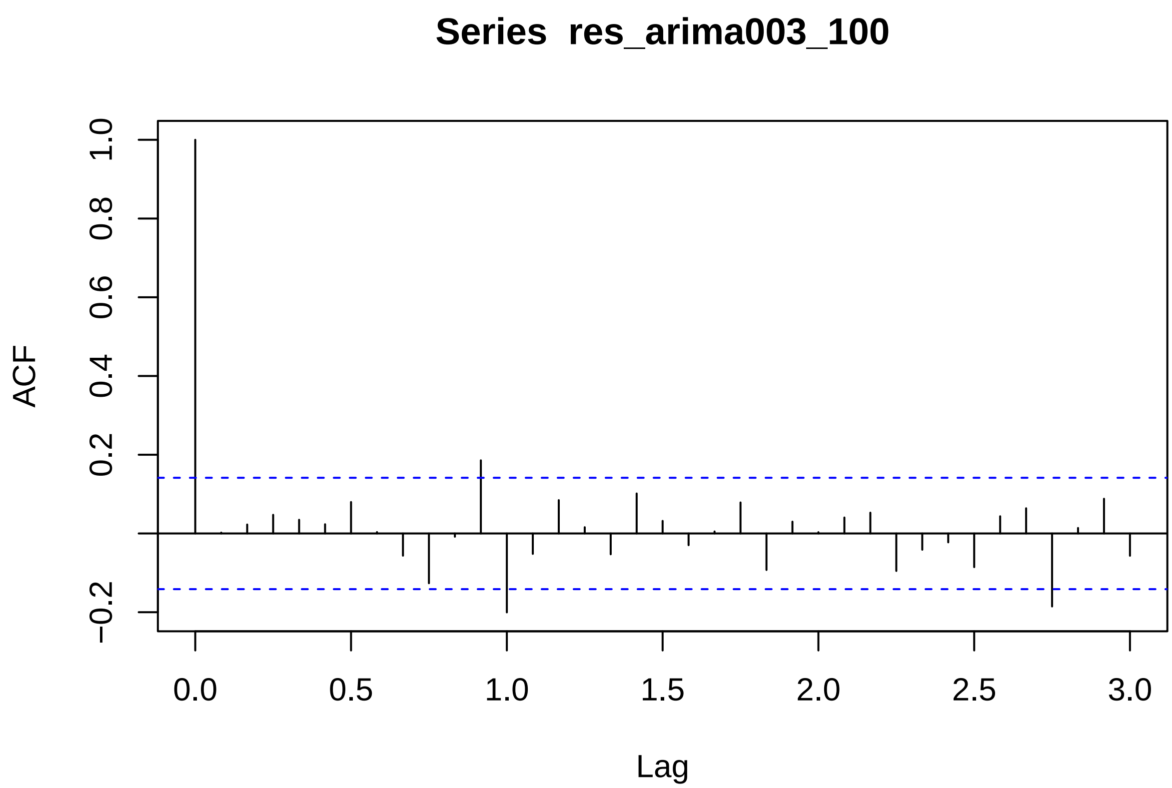
RMSE = 1250. The seasonal lags are still significant on the ACF plot. On the PACF plot, the first eight lags are not significant. However, lags (9,15,16) are significant. Along with lag 12, this still indicates that we should add a seasonal AR term on the undifferenced data. We do see a lower RMSE value on this model. A previous unsuccessful attempt at a ${ARIMA(0,0,2)} had RMSE = 1329.

ARIMA(0,0,3)(1,0,0) with constant and seasonal

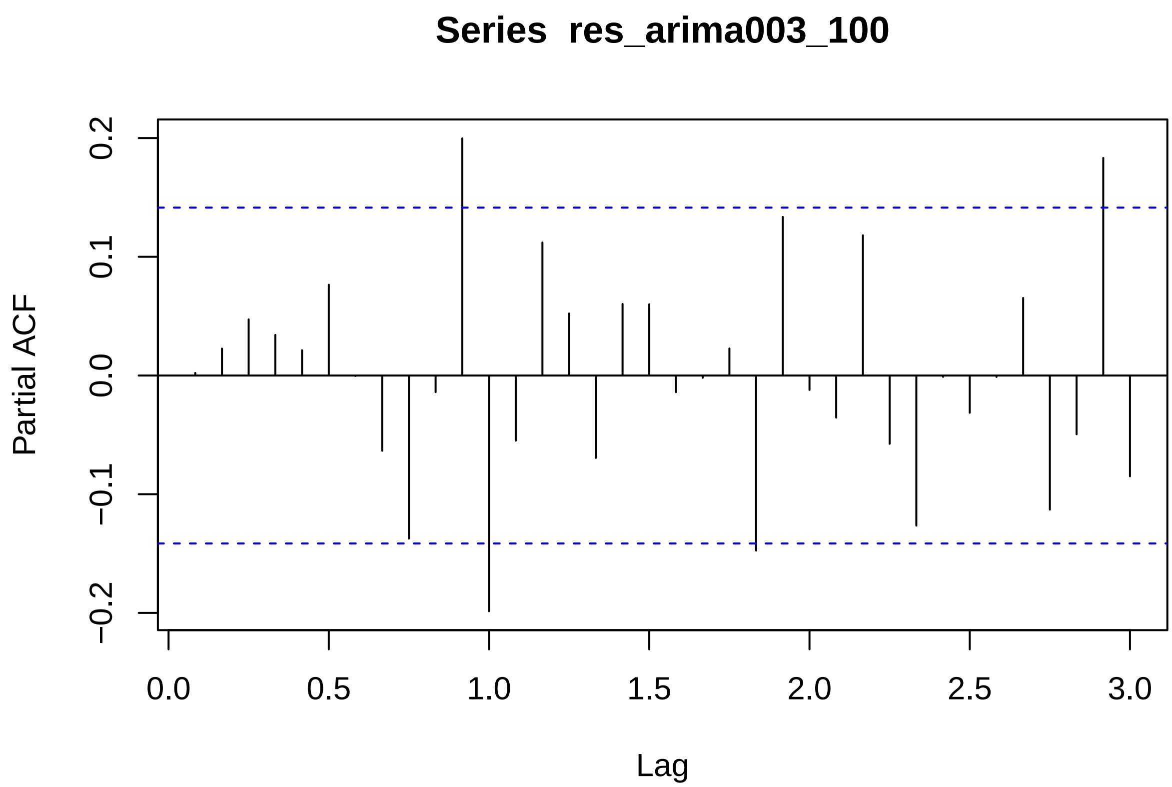
ARIMA(0,0,3)(1,0,0)$_{12}$

```
si_arima003_100_fit_cons <- si_ts2 %>%
  model(arima003_100_constant = ARIMA(total_waste ~ 1 +
                                        pdq(0,0,3) +
                                        PDQ(1,0,0, period = 12)))

zoo_arima003_100_fit <- arima(DSNY_SI_zoo_ts,
                              order = 1 + c(0,0,3),
                              seasonal = list(order = c(1, 0L, 0L),
```

```
                                                 period = 12))
res_arima003_100 <- zoo_arima003_100_fit$residuals
acf(res_arima003_100, lag.max = 36)
```




```
pacf(res_arima003_100, lag.max = 36)
```




```
accuracy(si_arima003_100_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 1388.
```

RMSE = 1387.547. The bad thing we see here is that the RMSE has increased.The majority of the lags in the ACF plot are contained within the bounds. The first ten lags in the PACF plot are contained within significant bounds, however, lags = 11, 12, 35 are not.

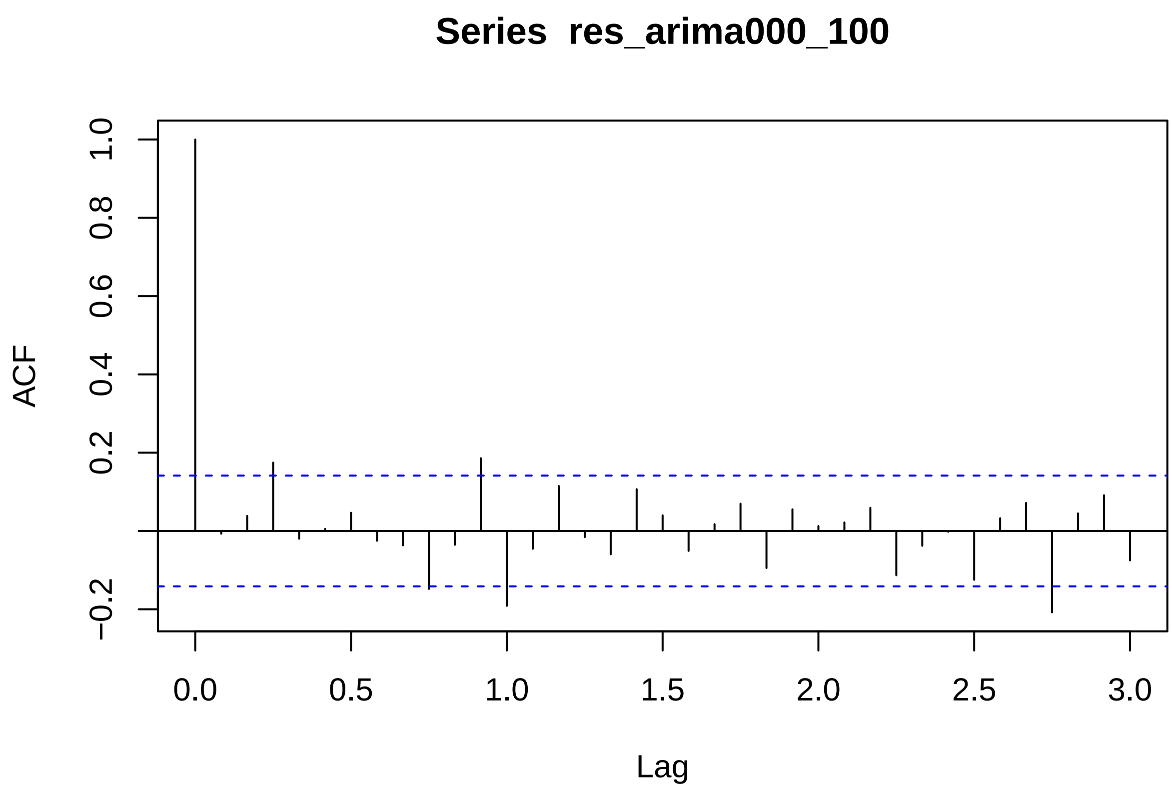
ARIMA(0,0,0)(1,0,0) with constant and seasonal

ARIMA(0,0,0)$(1,0,0)_{12}$

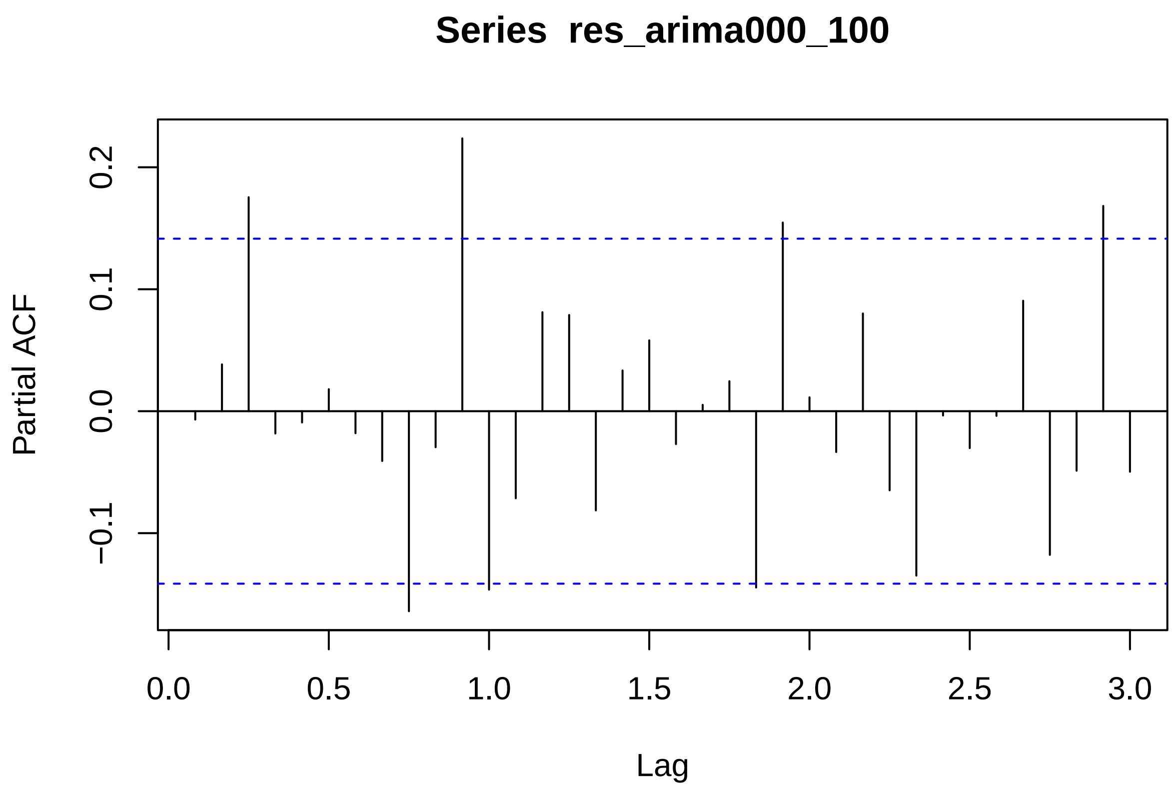
I am trying this model before moving onto the differenced data. This model worked for the BX tonnage values.

```
si_arima000_100_fit_cons <- si_ts2 %>%
  model(arima000_100_constant = ARIMA(total_waste ~ 1 +
                                        pdq(0,0,0) +
                                        PDQ(1,0,0, period = 12)))

zoo_arima000_100_fit <- arima(DSNY_SI_zoo_ts,
```

```
                          order = 1 + c(0,0,0),
                          seasonal = list(order = c(1, 0L, 0L),
                                          period = 12))
#names(zoo_arima000_fit)
res_arima000_100 <- zoo_arima000_100_fit$residuals
acf(res_arima000_100, lag.max = 36)
```




```
pacf(res_arima000_100, lag.max = 36)
```



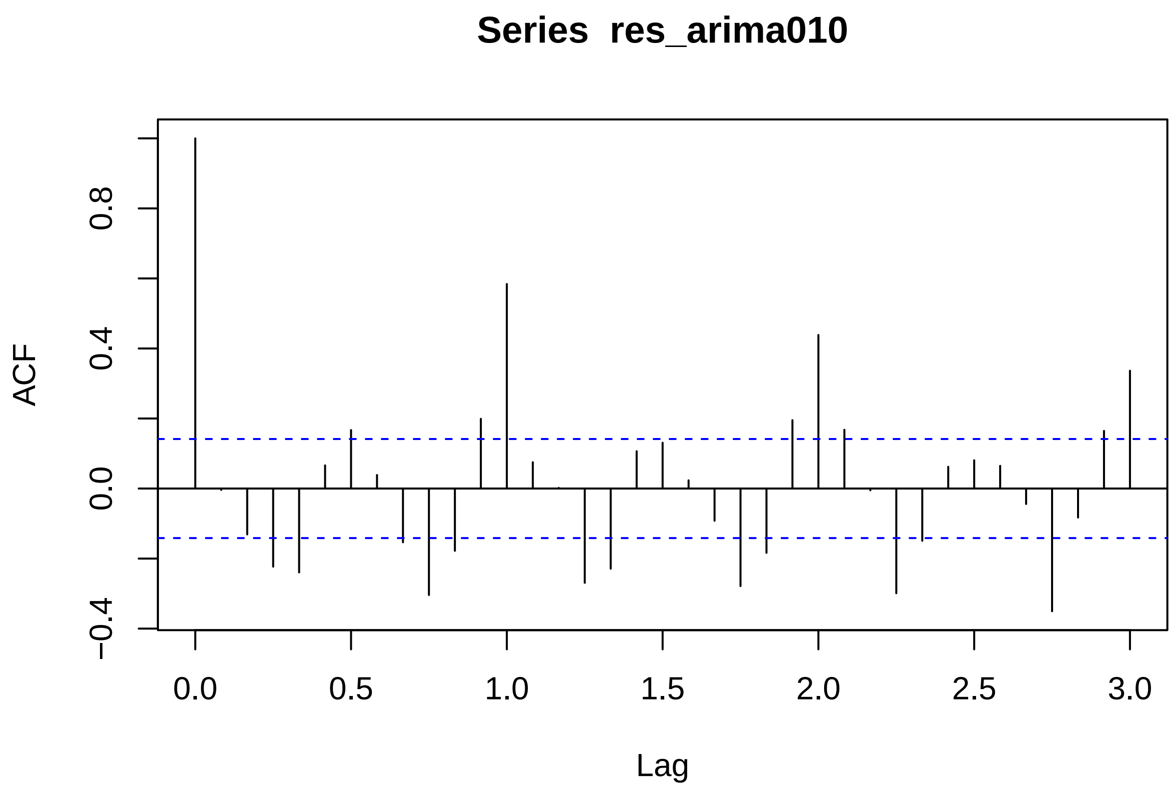

```
accuracy(si_arima000_100_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 1470.
```

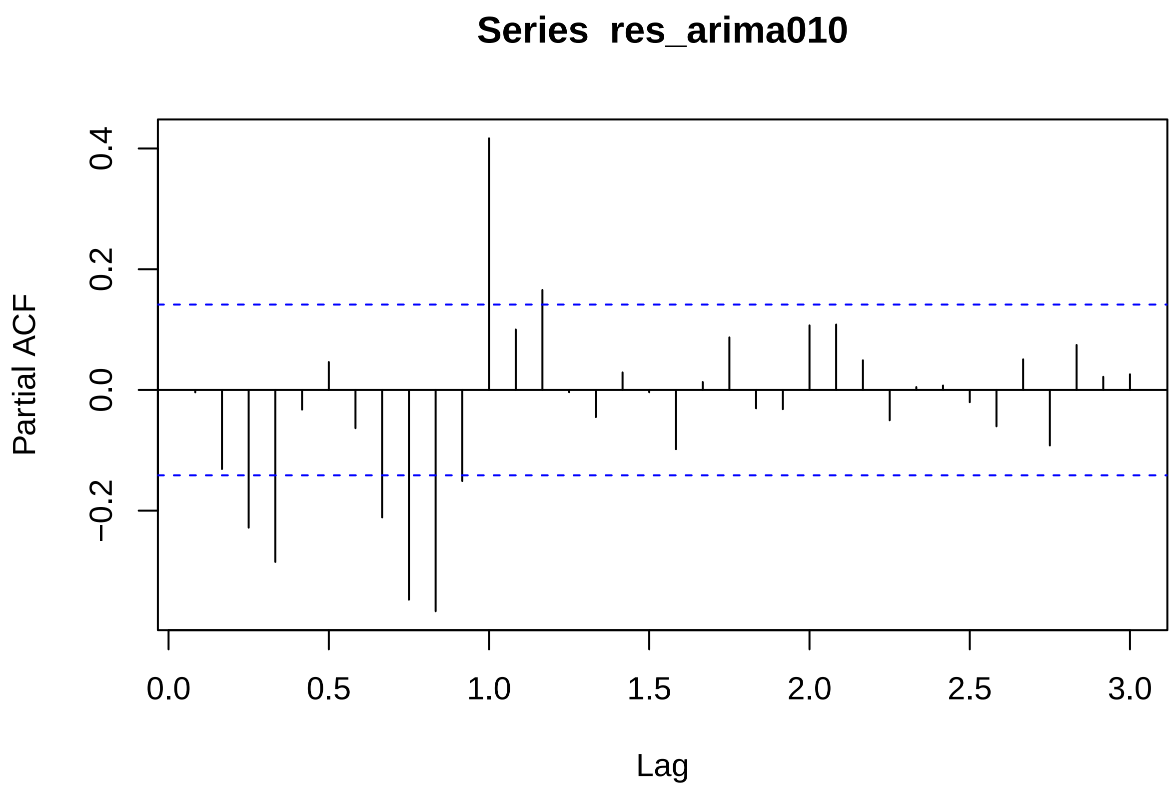
The RMSE continue to rise and the lags on the PACF plot do not improve. Onto the differenced data.

ARIMA(0,1,0) with constant

```
# si_ts2 %>%
#   model(arima010_constant = ARIMA(total_waste ~ pdq(0,1,0) + 1))

si_arima010_fit_cons <- si_ts3 %>%
  model(arima010_constant = ARIMA(total_waste ~ 1 + pdq(0,1,0)))

zoo_arima010_fit <- arima(DSNY_SI_zoo_ts,
                          order = 1 + c(0,1,0))
#names(zoo_arima000_fit)
res_arima010 <- zoo_arima010_fit$residuals
acf(res_arima010, lag.max = 36)
```




```
pacf(res_arima010, lag.max = 36)
```




```
accuracy(si_arima010_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 2165.
```

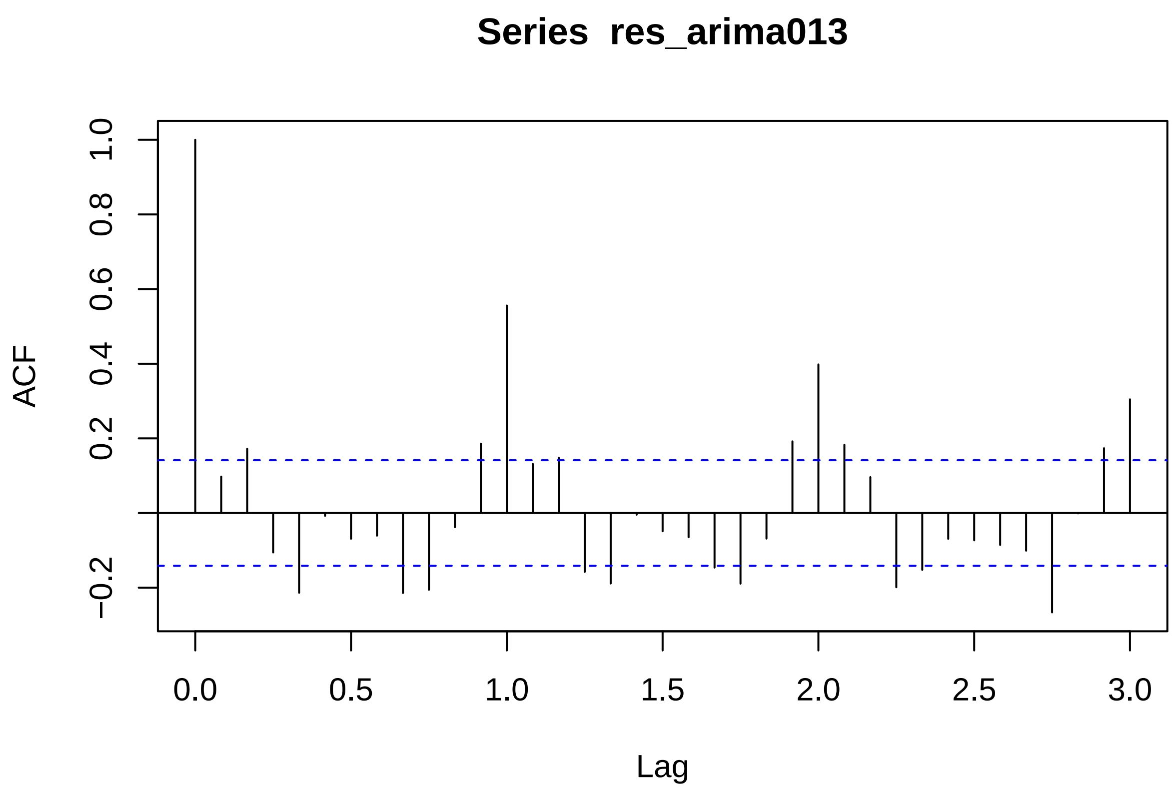
The PACF plot does show significant negative values at lags 3 and 4. And a positive significant lag at lag 12. We begin with an RMSE of 2164.96

ARIMA(0,1,3) with constant

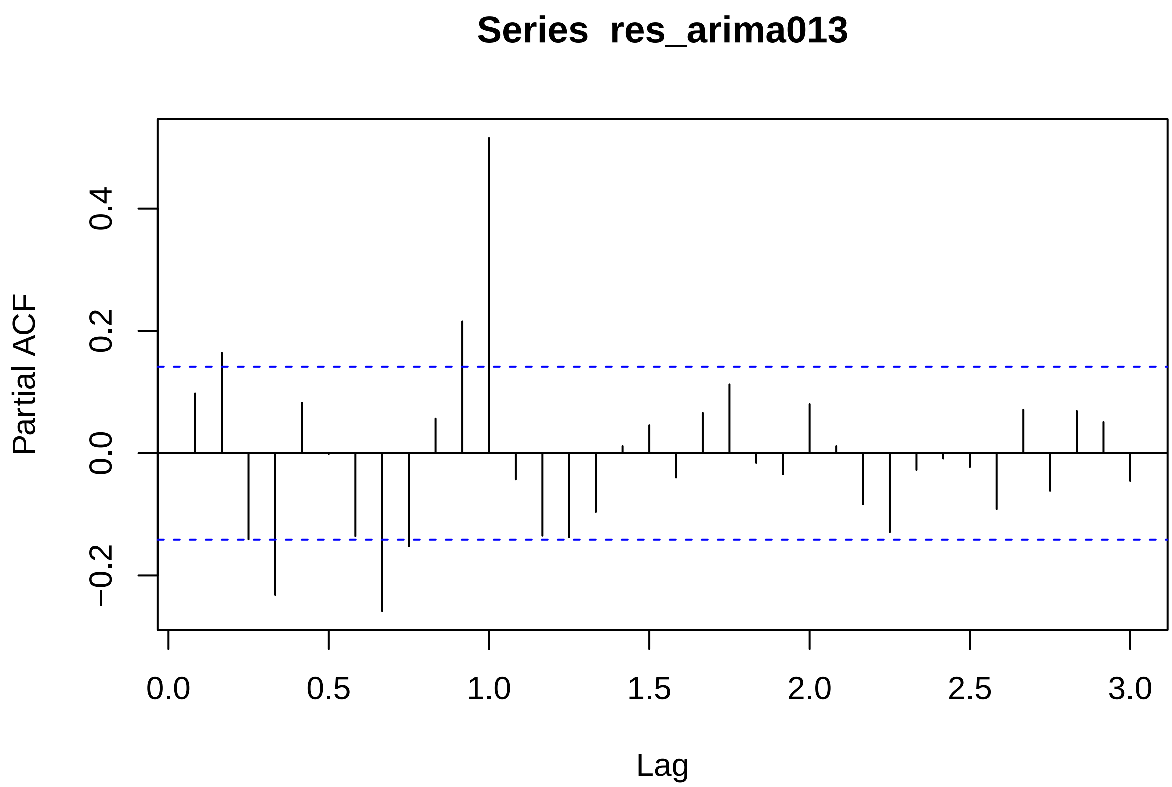
$ARIMA(0,1,3)$

```
# si_ts2 %>%
#   model(arima000_constant = ARIMA(total_waste ~ 1 +
#                                       pdq(0,0,0)))

si_arima013_fit_cons <- si_ts3 %>%
  model(arima013_constant = ARIMA(total_waste ~ 1 +
                                    pdq(0,1,3)))

zoo_arima013_fit <- arima(DSNY_SI_zoo_ts,
                          order = 1 + c(0,1,3))
```

```r
#names(zoo_arima000_fit)
res_arima013 <- zoo_arima013_fit$residuals
acf(res_arima013, lag.max = 36)
```




```r
pacf(res_arima013, lag.max = 36)
```



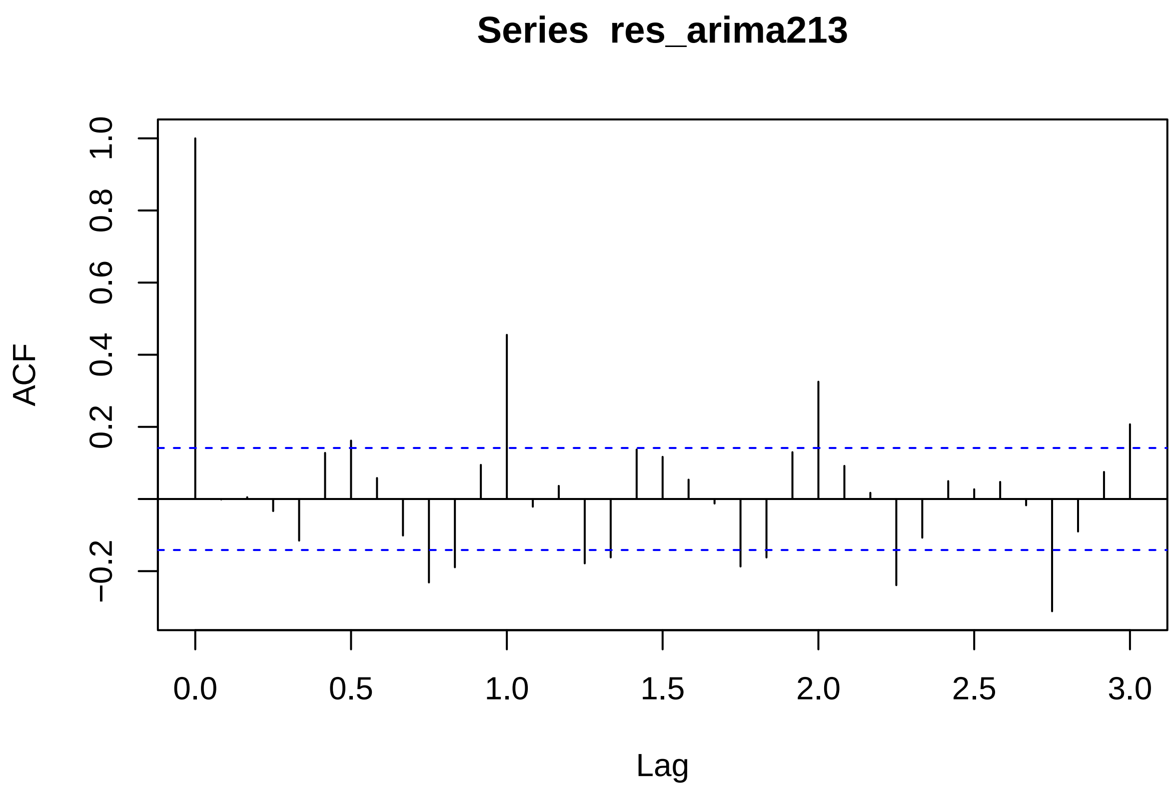

```
accuracy(si_arima013_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##     RMSE
##    <dbl>
## 1 1910.
```

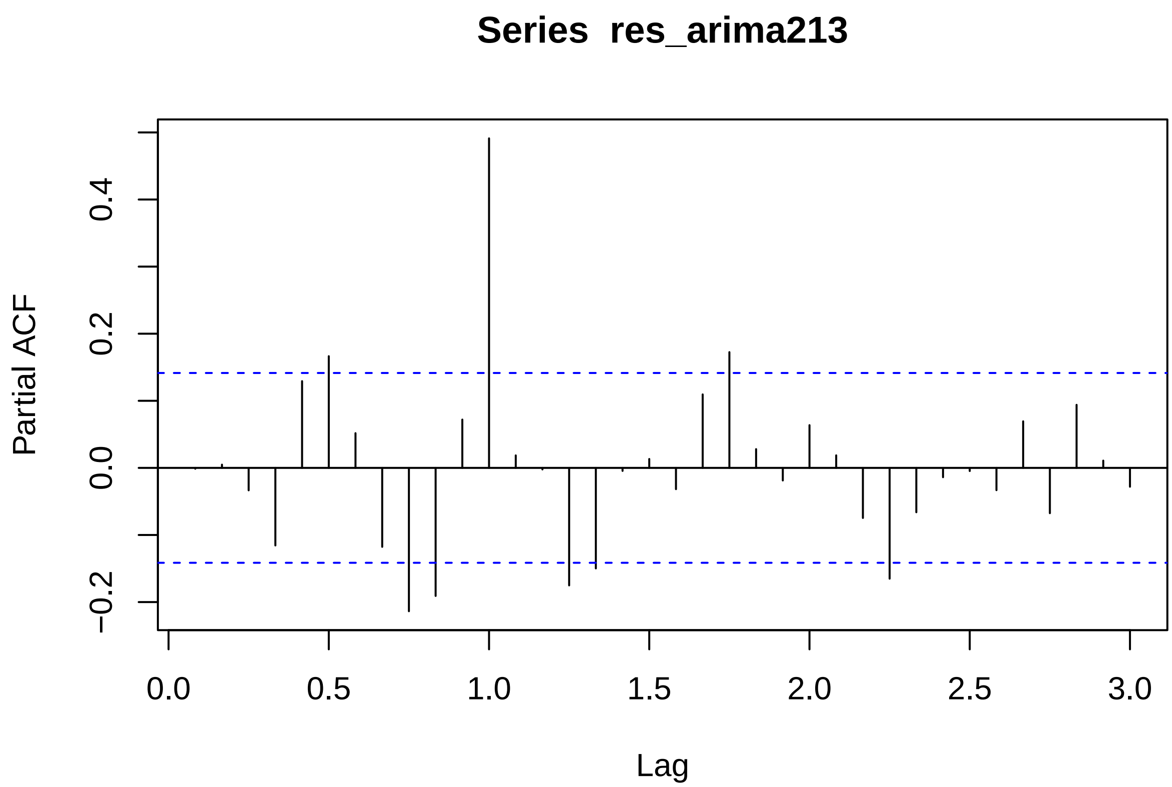
A good sign is that our RMSE value decreased. What I don't like to see is a positive significant value on the PACF at lag = 2. Adding a p = 2 would most likely delete any progress made. But let's attempt this first.

ARIMA(2,1,3) with constant

```
si_arima213_fit_cons <- si_ts3 %>%
  model(arima213_constant = ARIMA(total_waste ~ 1 +
                                    pdq(2,1,3)))

zoo_arima213_fit <- arima(DSNY_SI_zoo_ts,
                          order = 1 + c(2,1,3))
#names(zoo_arima000_fit)
res_arima213 <- zoo_arima213_fit$residuals
acf(res_arima213, lag.max = 36)
```




```
pacf(res_arima213, lag.max = 36)
```




```
accuracy(si_arima213_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 1800.
```

RMSE = 1800. In the ACF plot, the majority of the lags are not significant. However, the seasonal lags are significant and positive. In the PACF plot, lags 1-5 are not significant, but we see a significant and positive lag at 6, negative at lags (9,10). The seasonal lags spike again, so lets address that first.

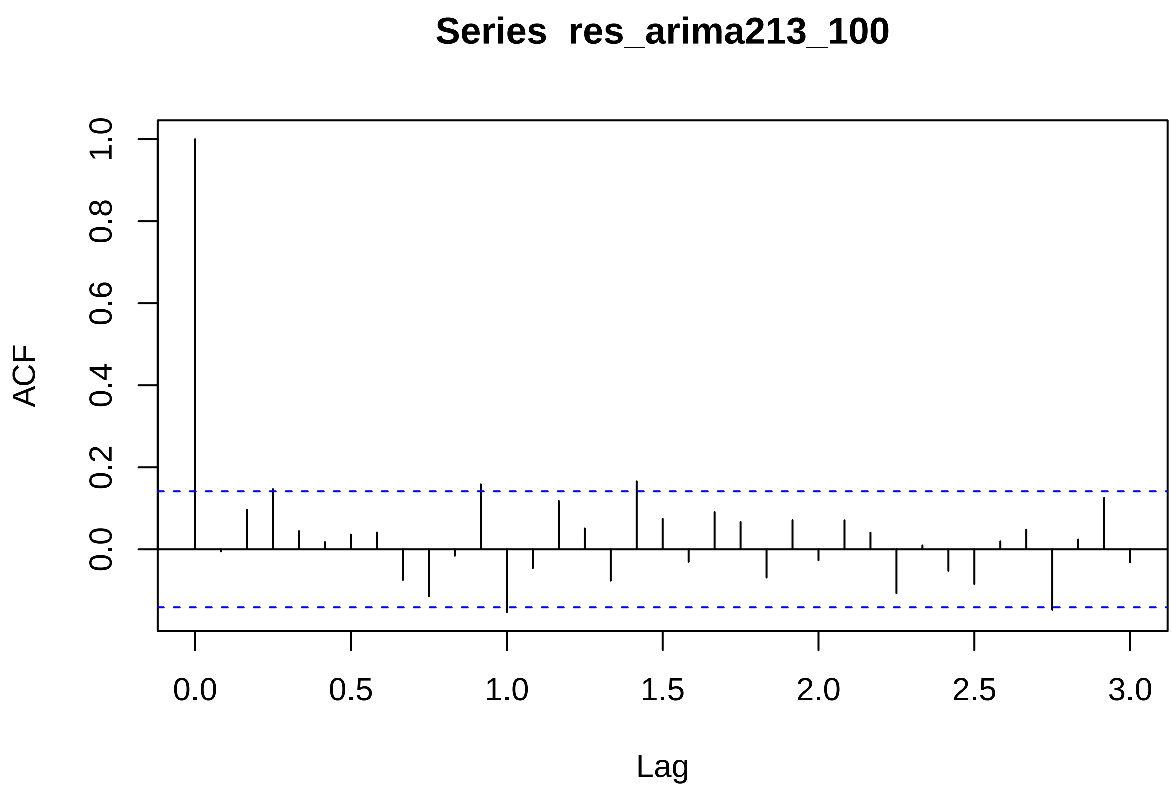
ARIMA(2,1,3)(1,0,0) with constant

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

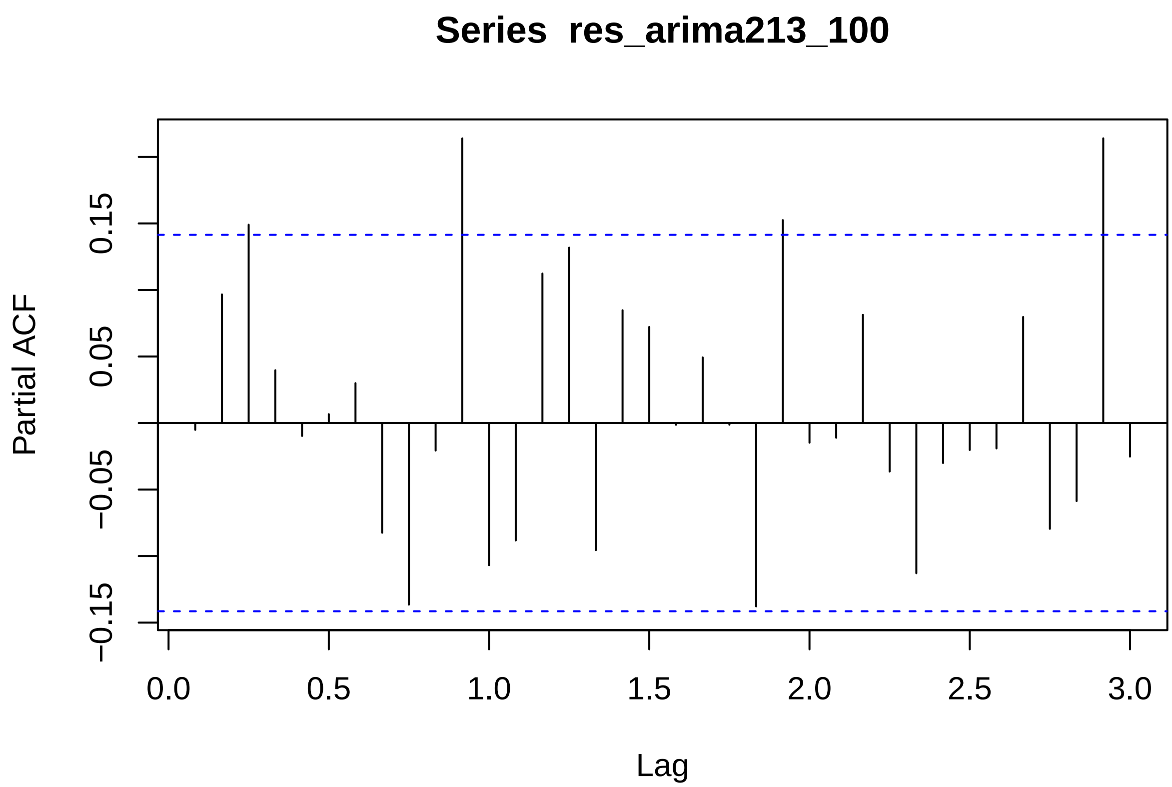
Lets deal with the common seasonality patter with this same arguments for now, to see if that can help stabilize the PACF values

```
si_arima213_100_fit_cons <- si_ts2 %>%
  model(arima213_100_constant = ARIMA(total_waste ~ 1 +
                                      pdq(2,1,3) +
                                      PDQ(1,0,0, period = 12)))

zoo_arima213_100_fit <- arima(DSNY_SI_zoo_ts,
```

```r
                                order = 1 + c(2,1,3),
                                seasonal = list(order = c(1, 0L, 0L),
                                                period = 12))
#names(zoo_arima000_fit)
res_arima213_100 <- zoo_arima213_100_fit$residuals
acf(res_arima213_100, lag.max = 36)
```




```r
pacf(res_arima213_100, lag.max = 36)
```




```
accuracy(si_arima213_100_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##     RMSE
##    <dbl>
## 1 1371.
```

RMSE = 1371. All of the lags in the ACF plot appear to be within bounds. The majority of the lags in the PACF plot are within the bounds. But lags = (11,23,35) are significant. The majority of the lags are bounded b/w (-0.15, 0.20). The RMSE score is good enough? Adding more parameters would perhaps create an overfitted model.

Auto-arima

```
si_auto_arima_fit_cons <- si_ts3 %>%
  model(stepwise = ARIMA(total_waste),
        search = ARIMA(total_waste,
                       stepwise = FALSE,
                       approximation = FALSE))
accuracy(si_auto_arima_fit_cons)[1:4]
```

```
## # A tibble: 2 x 4
```

```
##   .model   .type       ME  RMSE
##   <chr>    <chr>    <dbl> <dbl>
## 1 stepwise Training -42.2 1858.
## 2 search   Training -41.7 1746.
```

```
si_auto_arima_fit_cons %>% select(.model = stepwise) %>% report()
```

```
## Series: total_waste
## Model: ARIMA(4,1,1)
##
## Coefficients:
##          ar1      ar2      ar3      ar4      ma1
##       0.5881  -0.1243  -0.1567  -0.0872  -0.9207
## s.e.  0.0765   0.0834   0.0833   0.0759   0.0289
##
## sigma^2 estimated as 3562246:  log likelihood=-1710.15
## AIC=3432.3   AICc=3432.76   BIC=3451.81
```

```
si_auto_arima_fit_cons %>% select(.model = search) %>% report()
```

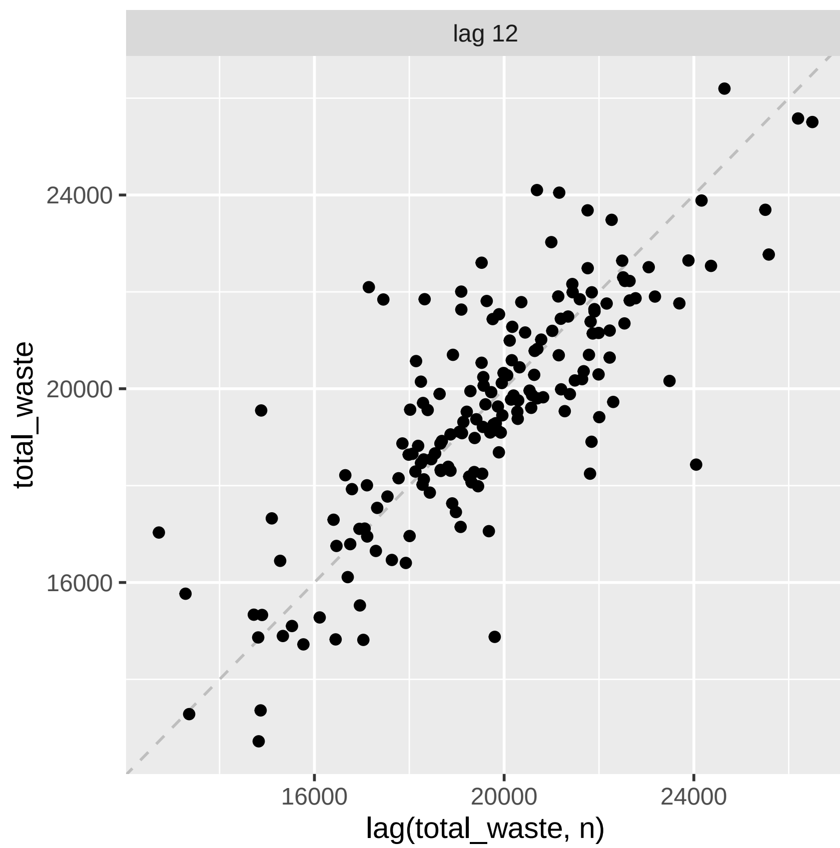
```
## Series: total_waste
## Model: ARIMA(2,1,4)
##
## Coefficients:
##         ar1      ar2      ma1     ma2      ma3      ma4
##       0.941  -0.9096  -1.3209  1.0062  -0.3099  -0.2705
## s.e.  0.059   0.0611   0.1004  0.1232   0.1245   0.0710
##
## sigma^2 estimated as 3164985:  log likelihood=-1698.89
## AIC=3411.77   AICc=3412.39   BIC=3434.54
```

Summary of the waste models

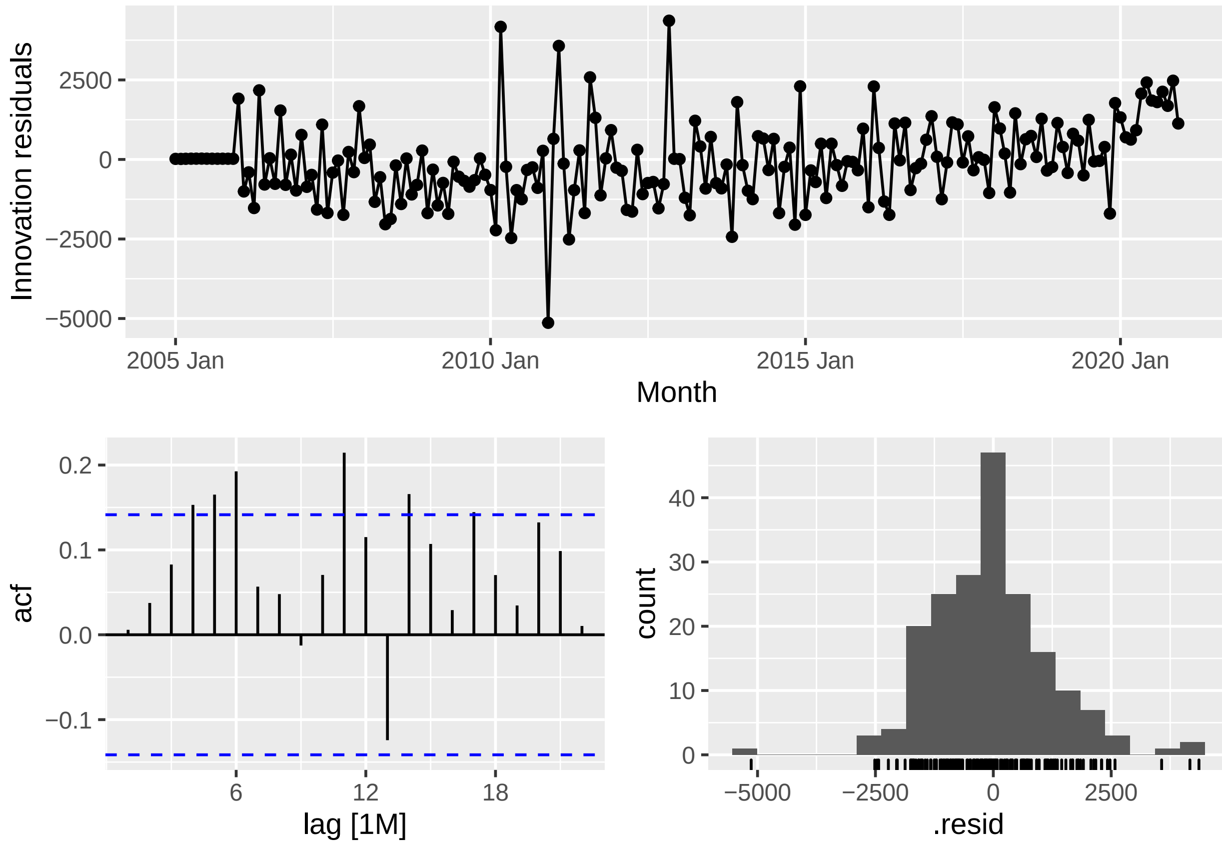
$ARIMA(0,0,3)$ with $RMSE \approx 1251$
$ARIMA(0,0,3)(1,0,0)_{12}$ with $RMSE \approx 1388$
$ARIMA(0,1,3)$ with $RMSE \approx 1910$
$ARIMA(2,1,3)(1,0,0)_{12}$ with $RMSE \approx 1371.112$ $ARIMA(2,1,4)$ with $RMSE \approx 1746.309$

Plots and visualizations

```
si_ts3 %>% gg_lag(total_waste, geom = "point", lags = 12)
```

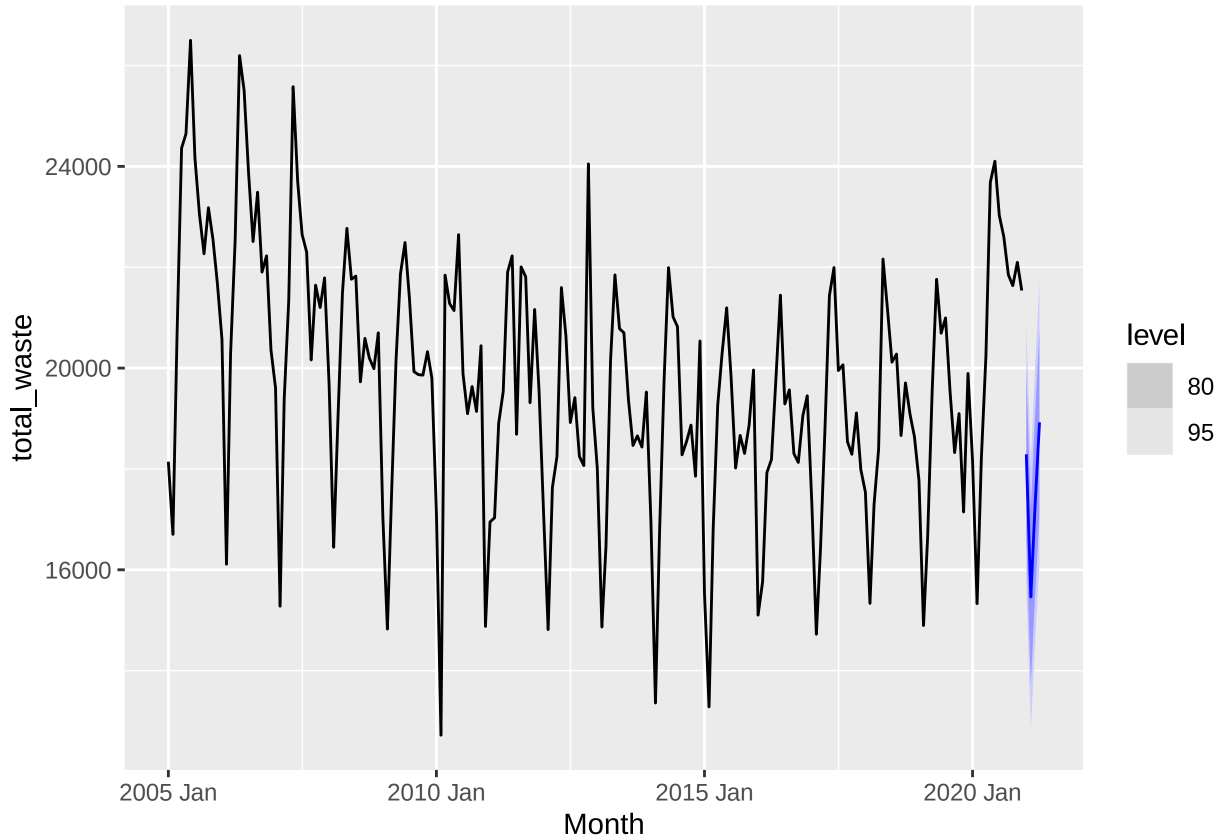



```
si_arima003_fit_cons %>% gg_tsresiduals()
```



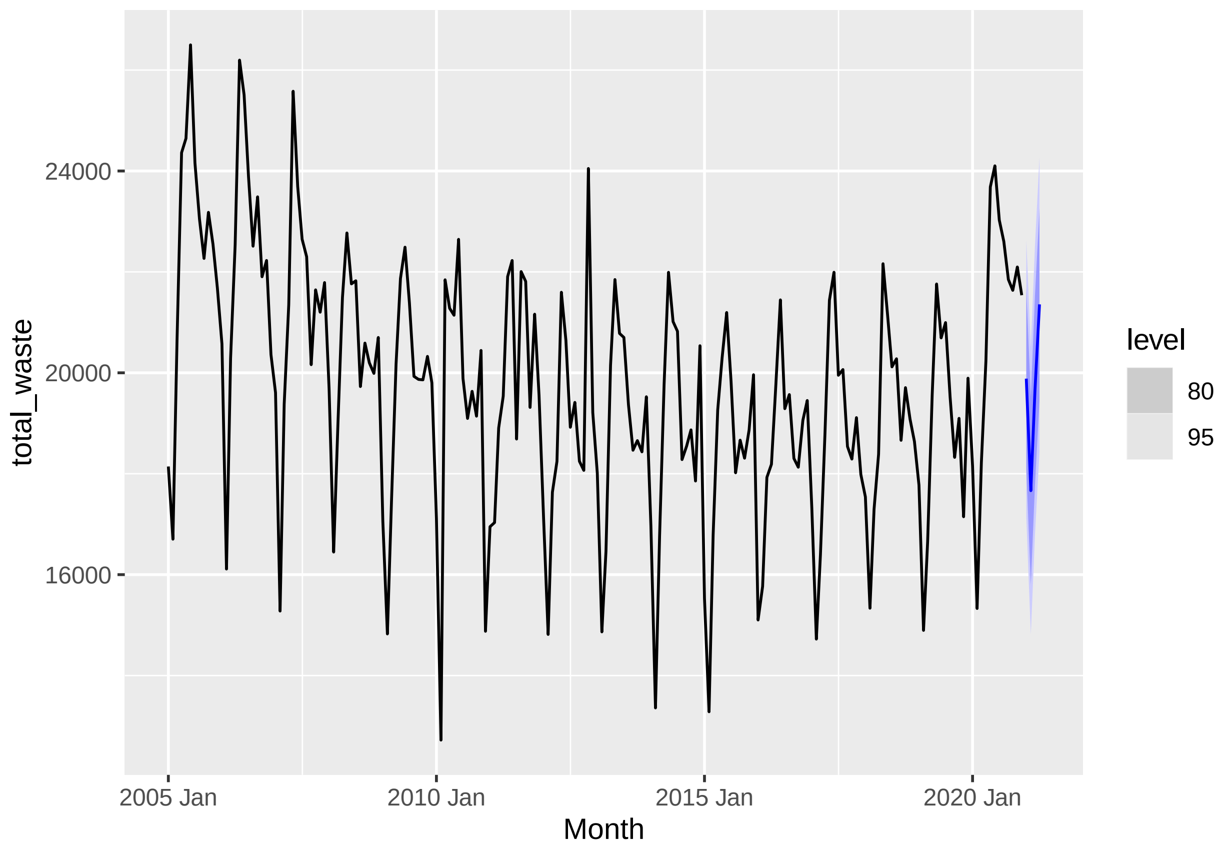
Preliminary forecast of si_arima002_fit_cons

```
si_arima003_fit_cons %>%
  forecast(h = 4) %>%
  autoplot(si_ts2)
```



Preliminary forecast of si_arima213_100_fit_cons

```
si_arima213_100_fit_cons %>% forecast(h = 4) %>% autoplot(si_ts2)
```



Refer to data_prep file, around line 557 and nineth_meeting_notes file around line 109 for more plots designs