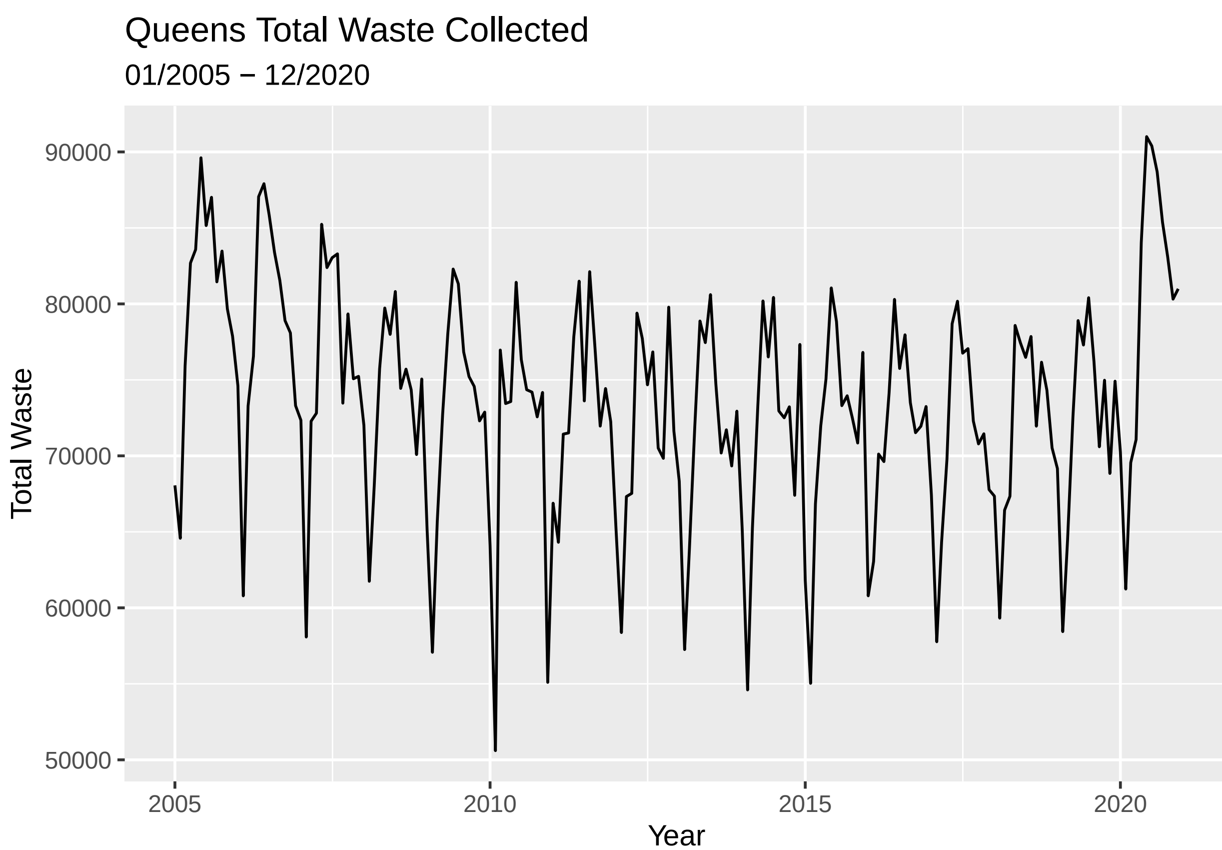
Queens Time Series

Daniel L.

5/4/2022

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




```
DSNY_third_queens %>%
  group_by("Year" = year(month)) %>%
  summarise(.,
            "Average Total Waste" = mean(total_waste))
```

```
## # A tibble: 16 x 2
##     Year `Average Total Waste`
##    <dbl>                 <dbl>
##  1  2005                79922.
##  2  2006                78440.
##  3  2007                76049.
##  4  2008                73809.
##  5  2009                72840.
##  6  2010                70569.
##  7  2011                73744.
##  8  2012                71562.
##  9  2013                71437.
## 10  2014                71563.
## 11  2015                71499.
## 12  2016                71832.
## 13  2017                71189.
## 14  2018                71980.
## 15  2019                72232.
## 16  2020                79668.
```

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.

```
#total waste values
qns_ts %>% features(total_waste, unitroot_kpss)
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1     0.346         0.1
```

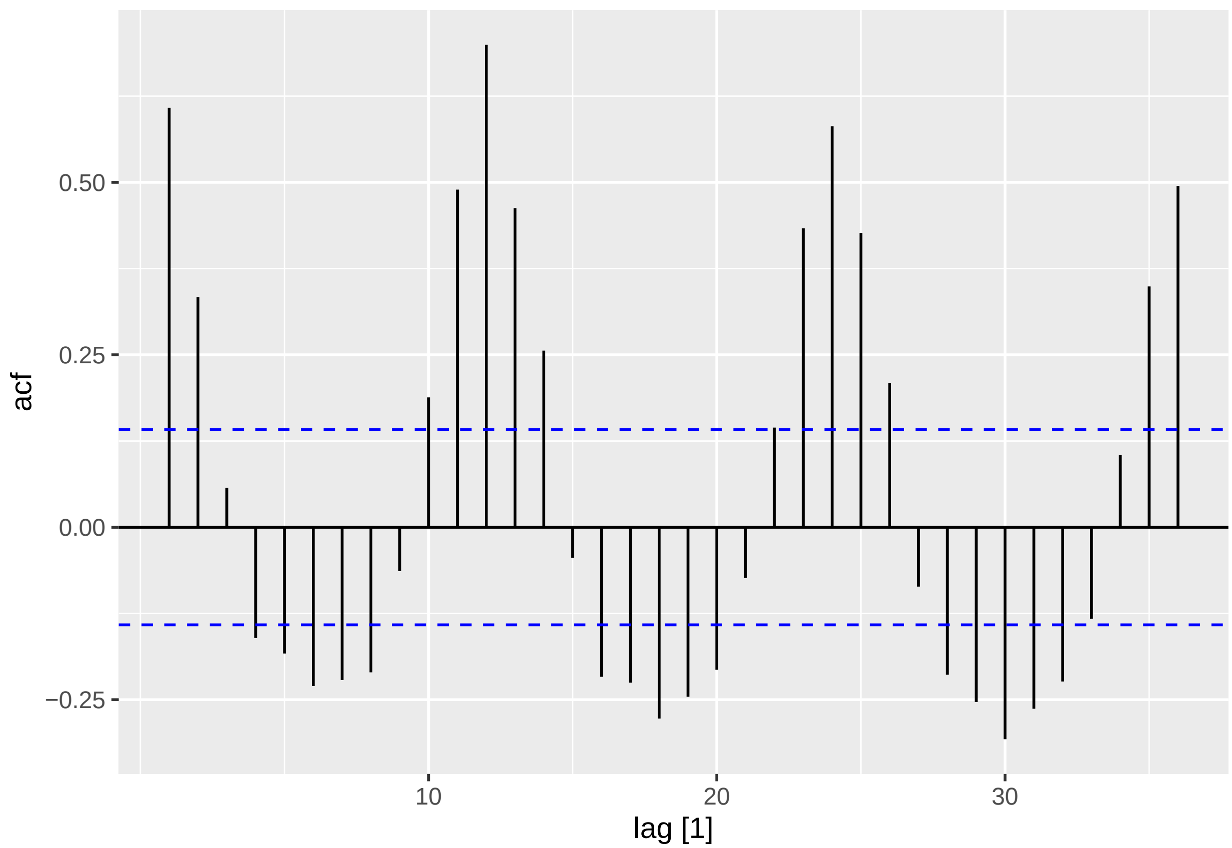
```
#differenced values
qns_ts %>% features(diff1, unitroot_kpss)
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1    0.0156         0.1
```

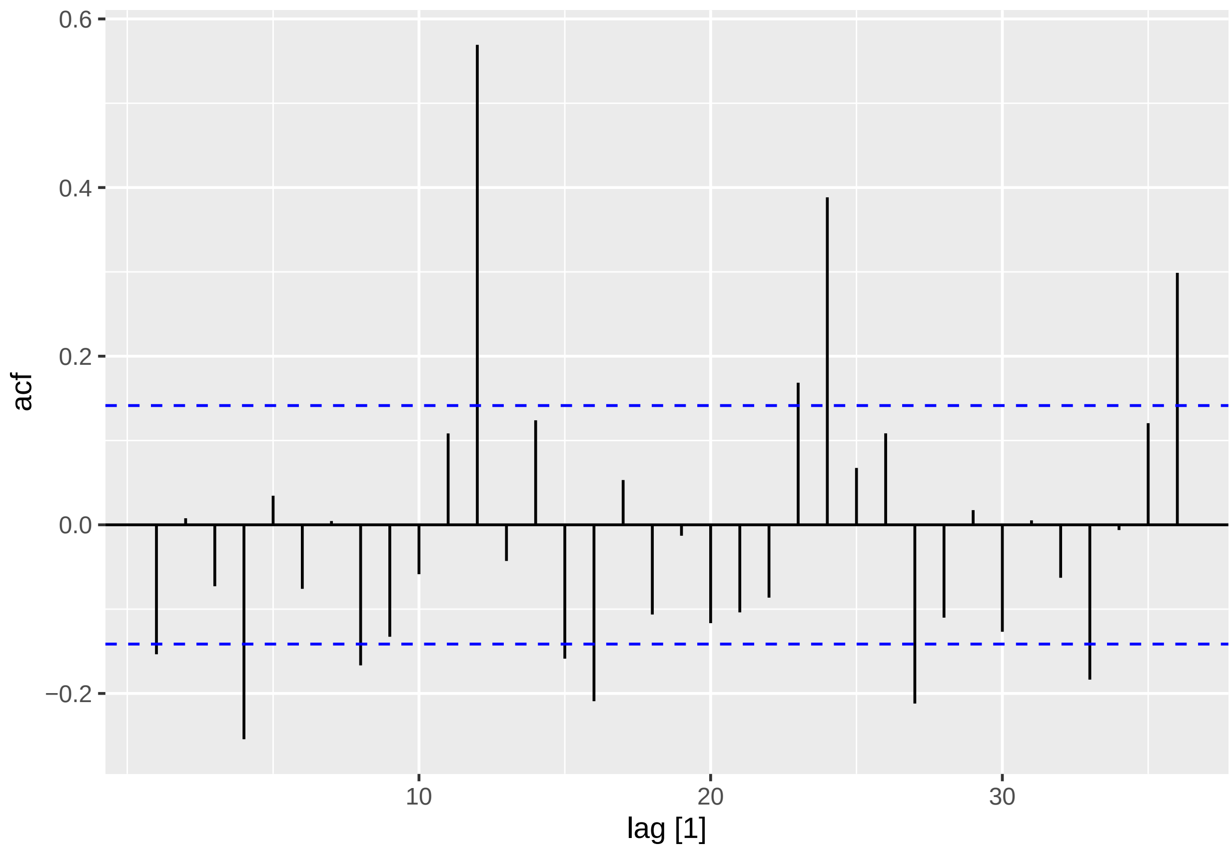
According the KPSS test, both the QNS total_waste and differenced values fail to reject the H_0.

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

```
qns_ts3 %>%
  ACF(total_waste, lag_max = 36) %>%
  autoplot()
```

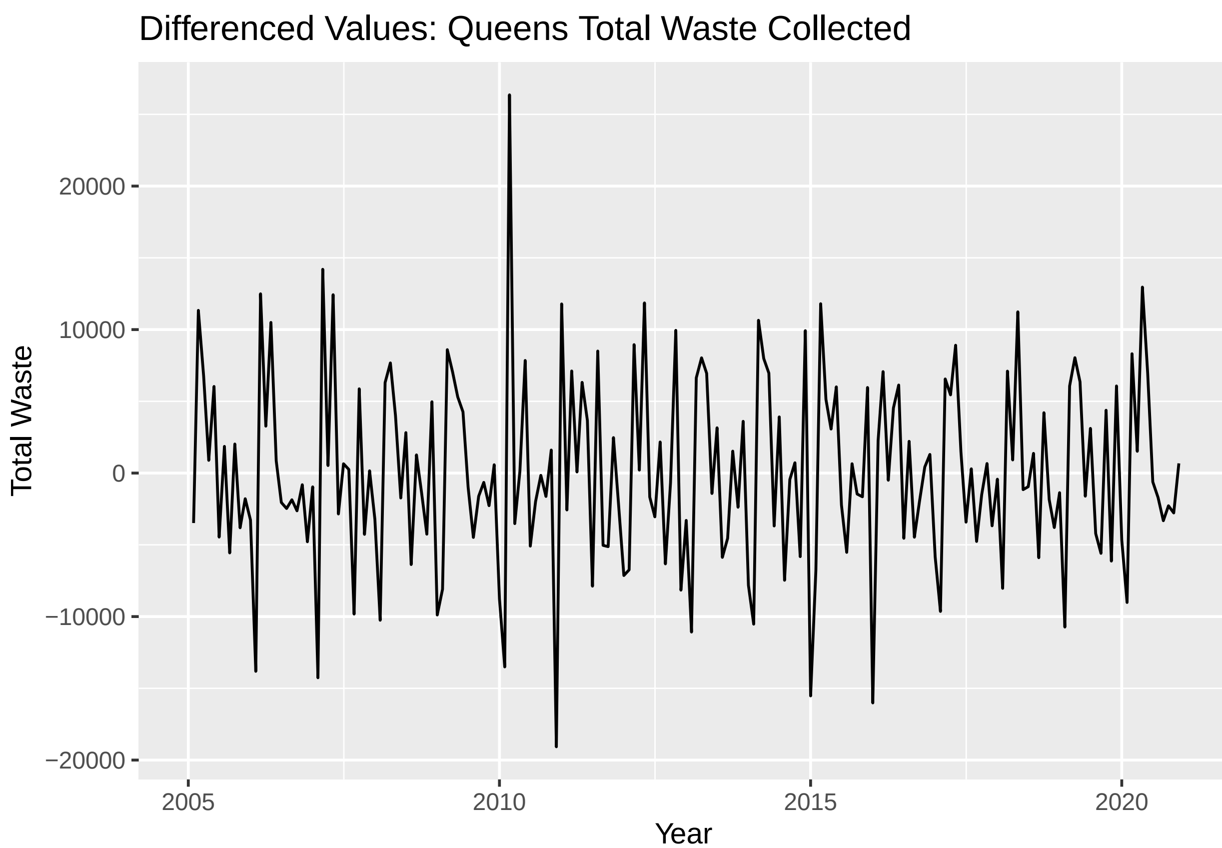



```
#acf of the differenced values
qns_ts3 %>%
  ACF(diff1, lag_max = 36) %>%
  autoplot()
```




```
qns_ts3 %>%
  ggplot(mapping = aes(x = month, y = diff1)) + geom_line() +
  labs(x = "Year",
       y = "Total Waste",
       title = "Differenced Values: Queens Total Waste Collected")
```

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



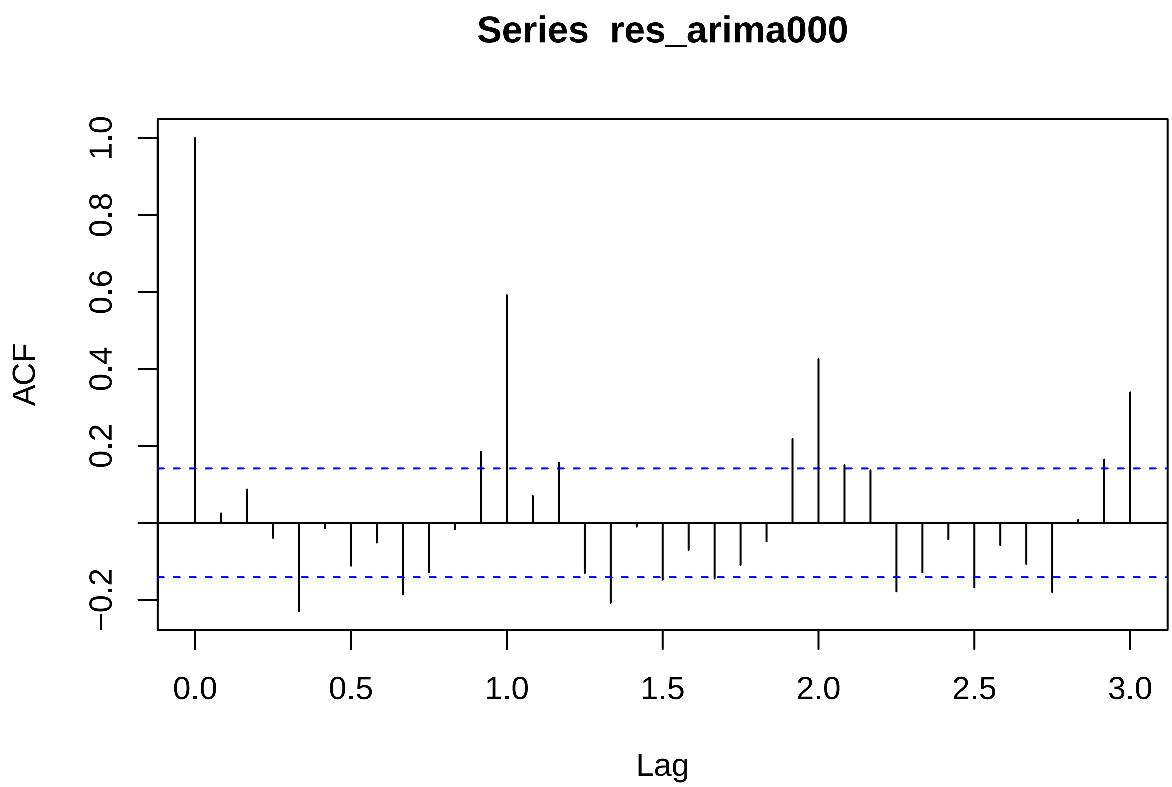
Creating models with zoo() and the arima package from stats()

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

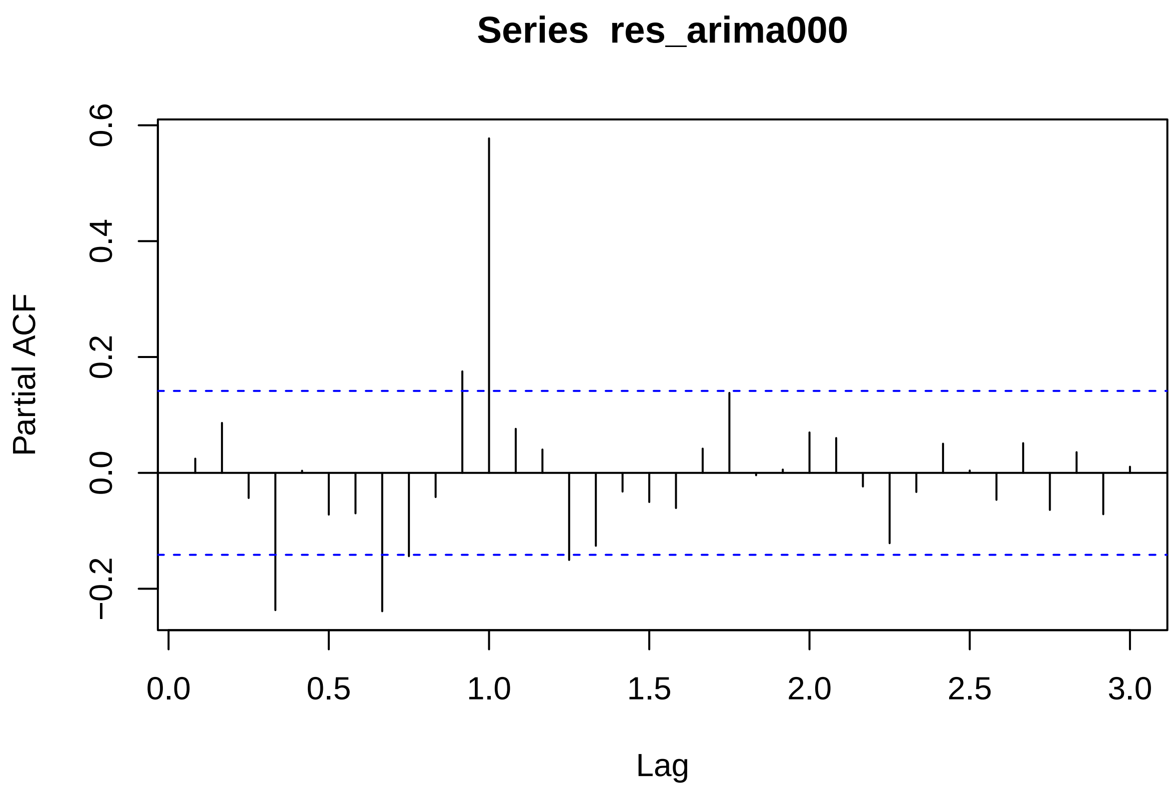
ARIMA(0,0,0) with constant

$ARIMA(0, 0, 0)$

```
qns_arima000_fit_cons <- qns_ts3 %>%
  model(arima000_constant = ARIMA(total_waste ~ 1 + pdq(0,0,0)))

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




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




```
accuracy(qns_arima000_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 7383.
```

RMSE = 7382.682. In the ACF and PACF, the first significant lag is lag = 4 and has a negative autocorrelation value.

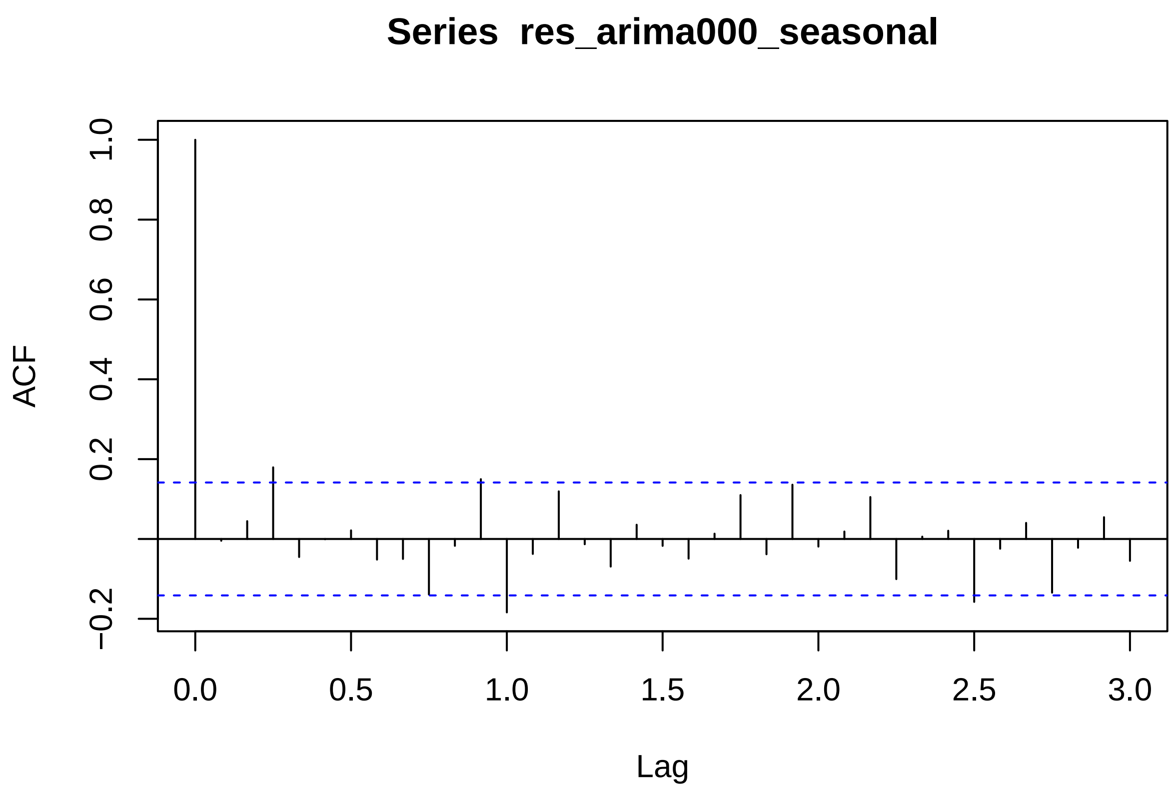
Before I add MA(4), I would like to see how the plots of a model with a seasonal AR(1)

ARIMA(0,0,0)(1,0,0)[12] with constant

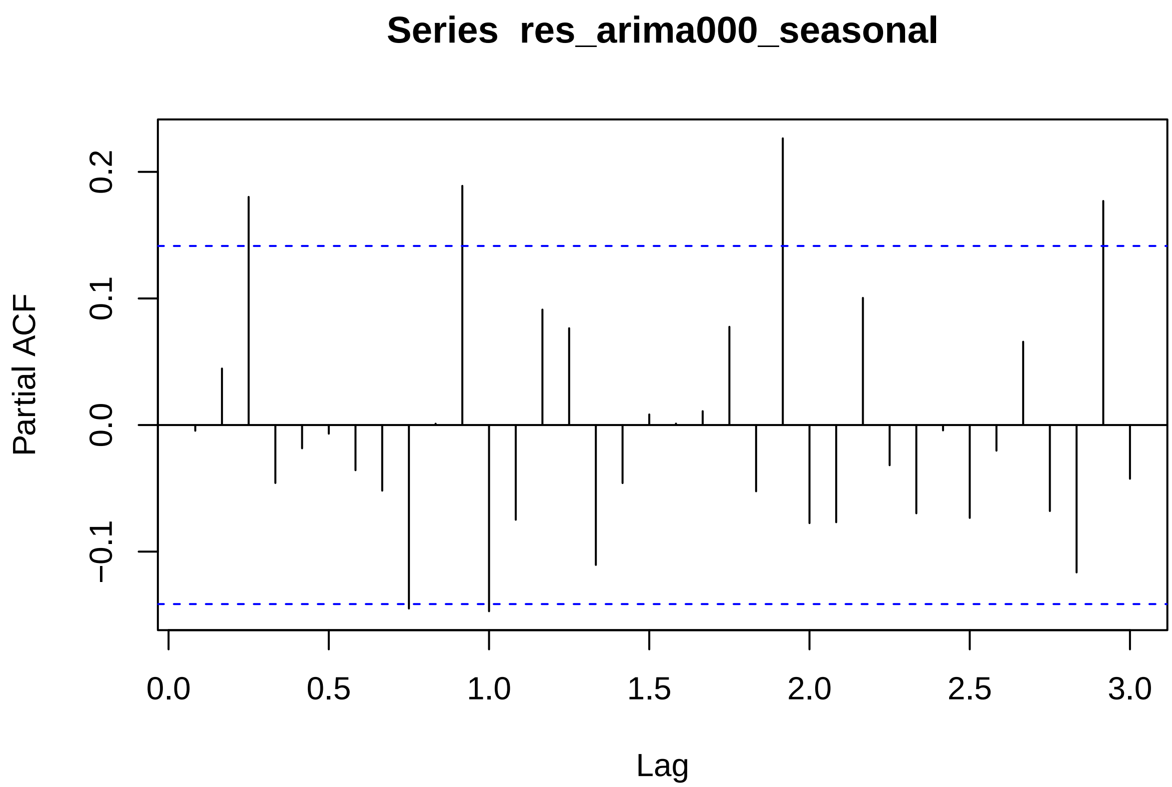
$ARIMA(0,0,0)(1,0,0)[12]$

```
qns_arima000_fit_seasonal_cons <- qns_ts3 %>%
  model(arima000_constant_seasonal = ARIMA(total_waste ~ 1 +
                                           pdq(0,0,0) +
                                           PDQ(1,0,0, period = 12)))

zoo_arima000_seasonal_fit <- arima(DSNY_QNS_zoo_ts,
                          order = 1 + c(0,0,0),
                          seasonal = list(order = c(1,0L,0L), period = 12))
```

```r
#names(zoo_arima000_fit)
res_arima000_seasonal <- zoo_arima000_seasonal_fit$residuals
acf(res_arima000_seasonal, lag.max = 36)
```




```r
pacf(res_arima000_seasonal, lag.max = 36)
```




```
accuracy(qns_arima000_fit_seasonal_cons)[4]
```

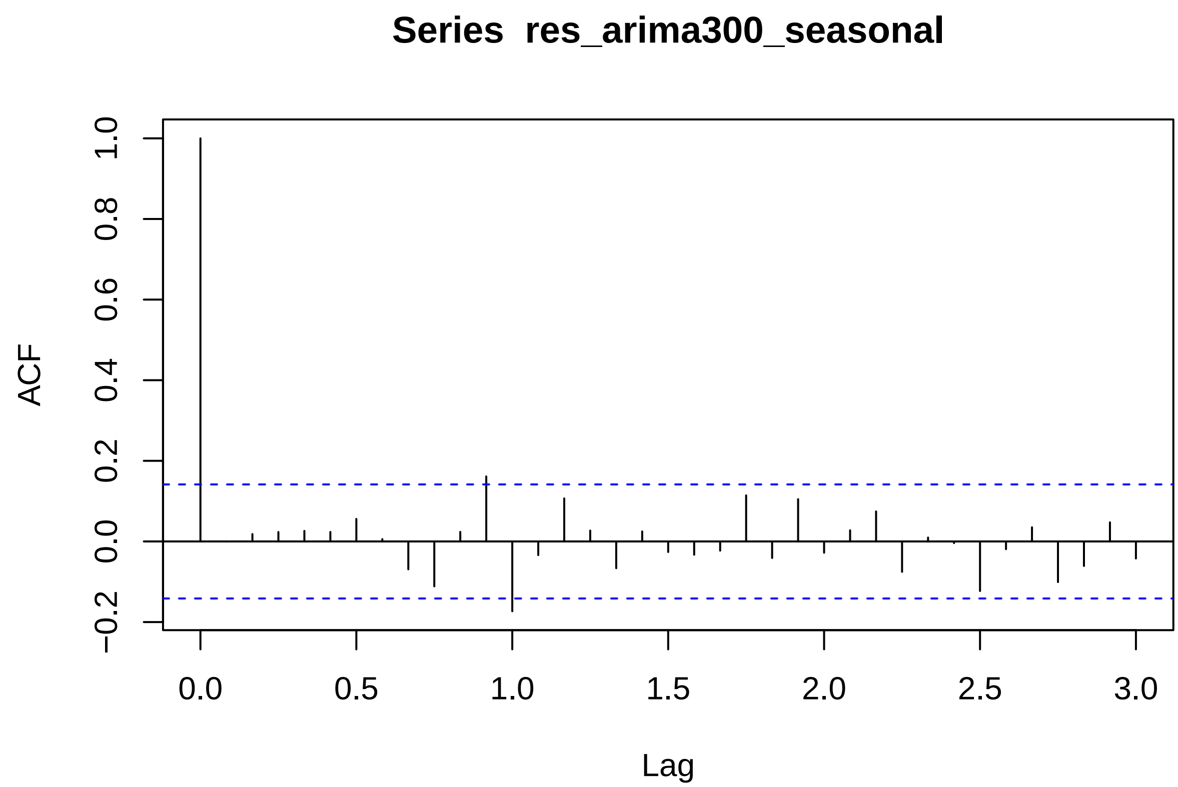
```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 4414.
```

RMSE = 4413.677. The seasonal lags appear to be contained. The PACF plot shows us that lag = 3 has positive correlation. Lags = (11,23,35) also have positive correlation.

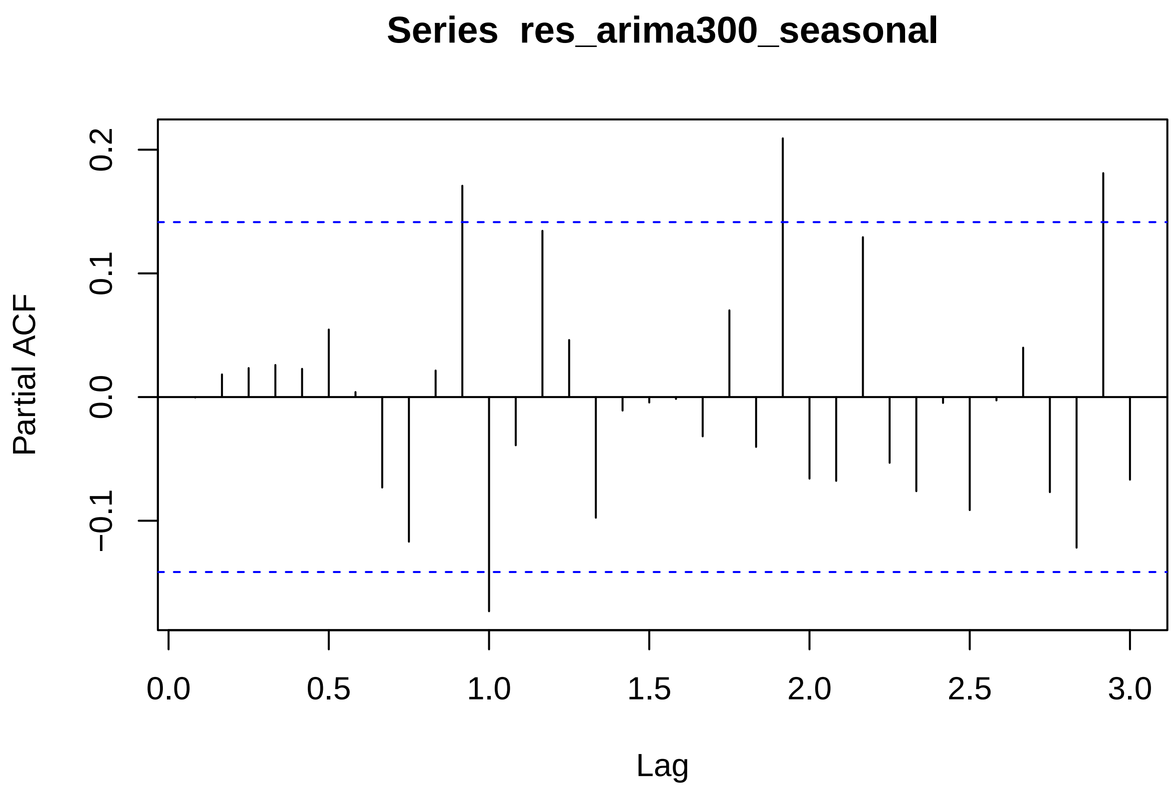
ARIMA(3,0,0)(1,0,0)[12] with constant

$ARIMA(3, 0, 0)(1, 0, 0)[12]$

```
qns_arima300_fit_seasonal_cons <- qns_ts3 %>%
  model(arima300_constant_seasonal = ARIMA(total_waste ~ 1 +
                                   pdq(3,0,0) +
                                   PDQ(1,0,0, period = 12)))

zoo_arima300_seasonal_fit <- arima(DSNY_QNS_zoo_ts,
                          order = 1 + c(3,0,0),
                          seasonal = list(order = c(1,0L,0L), period = 12))
#names(zoo_arima000_fit)
res_arima300_seasonal <- zoo_arima300_seasonal_fit$residuals
acf(res_arima300_seasonal, lag.max = 36)
```




```
pacf(res_arima300_seasonal, lag.max = 36)
```




```
accuracy(qns_arima300_fit_seasonal_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 4115.
```

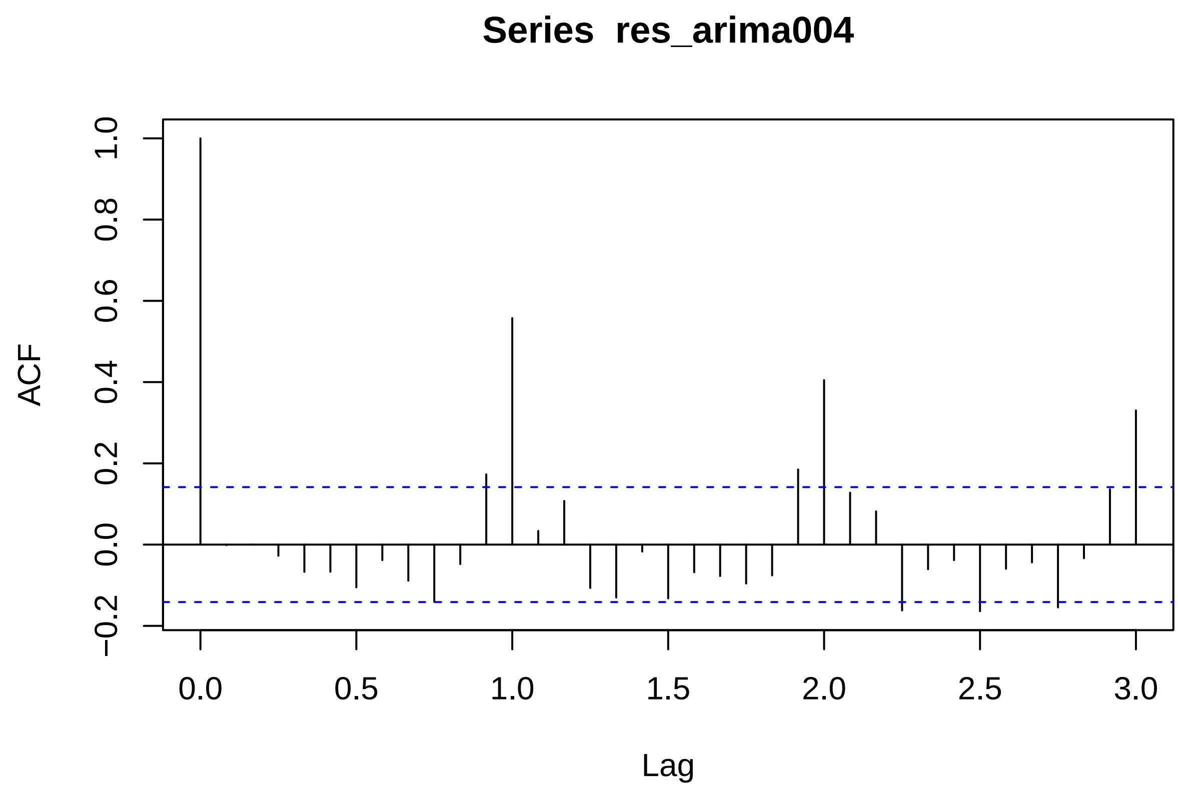
RMSE = 4115.02. We do see a decrease in the RMSE. The first 11 lags in the PACF have low autocorrelations. Lag = 12 has significant negative autocorrelation. Lags = (11, 23, 35) continue to be significant.

Lets work with an MA(4) model, with no seasonality.

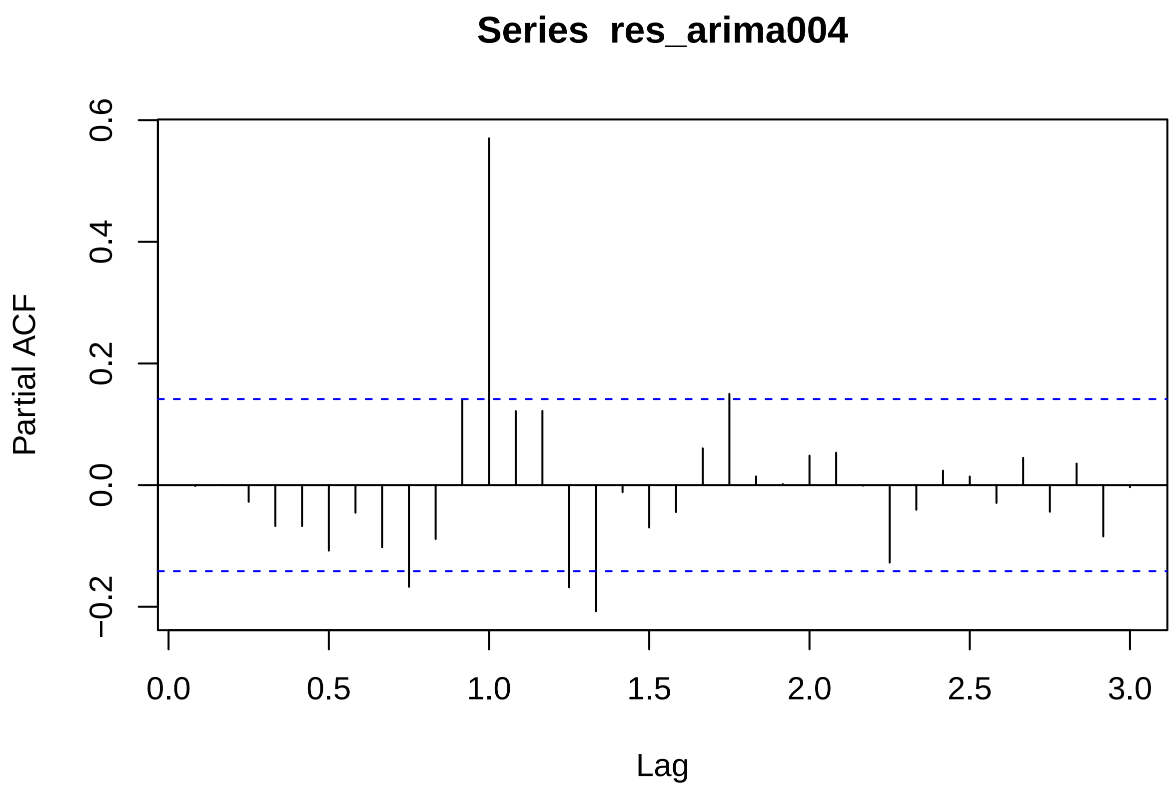
ARIMA(0,0,4) with constant

$ARIMA(0, 0, 4)$

```
qns_arima004_fit_cons <- qns_ts3 %>%
  model(arima004_constant = ARIMA(total_waste ~ 1 + pdq(0,0,4)))

zoo_arima004_fit <- arima(DSNY_QNS_zoo_ts, order = 1 + c(0,0,4))
res_arima004 <- zoo_arima004_fit$residuals
acf(res_arima004, lag.max = 36)
```




```
pacf(res_arima004, lag.max = 36)
```




```r
accuracy(qns_arima004_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 5658.
```

Our RMSE = 5657.673 is a decrease from the first model. Most of the autocorrelations in the ACF plot appear to be contained The first significant lag in the PACF plot is lag = 9, and has negative autocorrelation. Lag = 12 is very positively auto correlated.

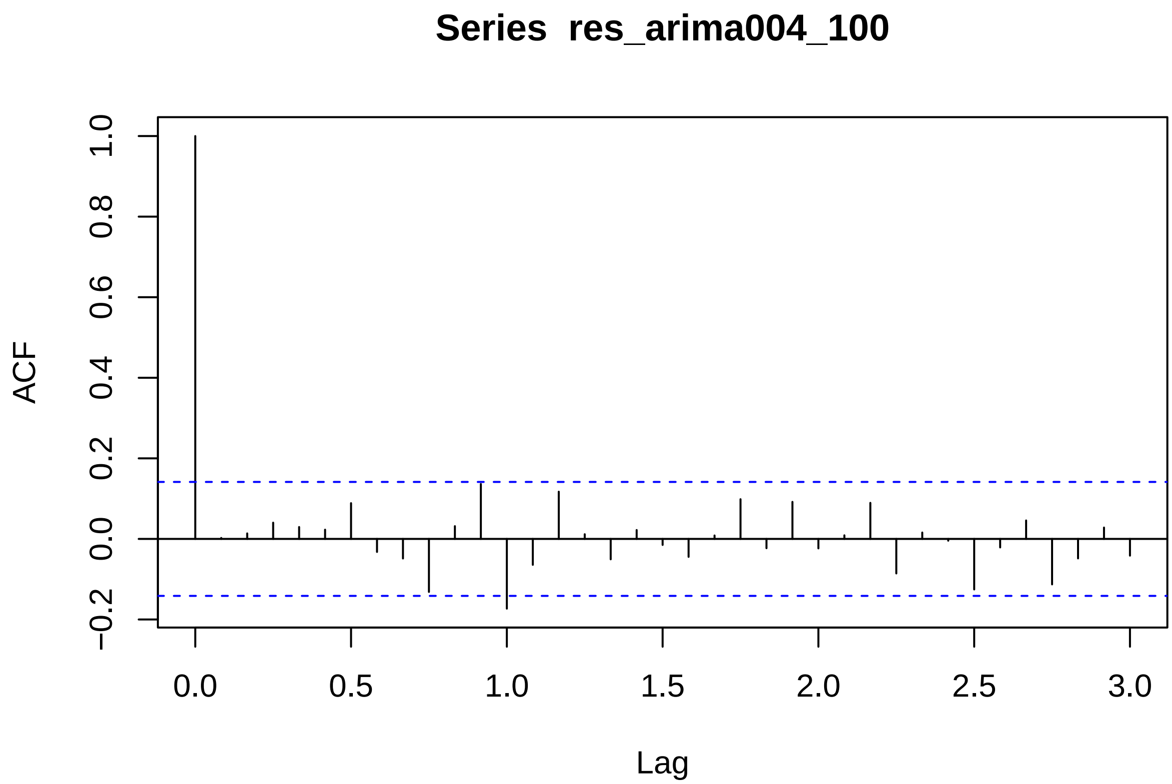
Lets work with an MA(4) model, with seasonality.

ARIMA(0,0,4)(1,0,0)[12] with constant

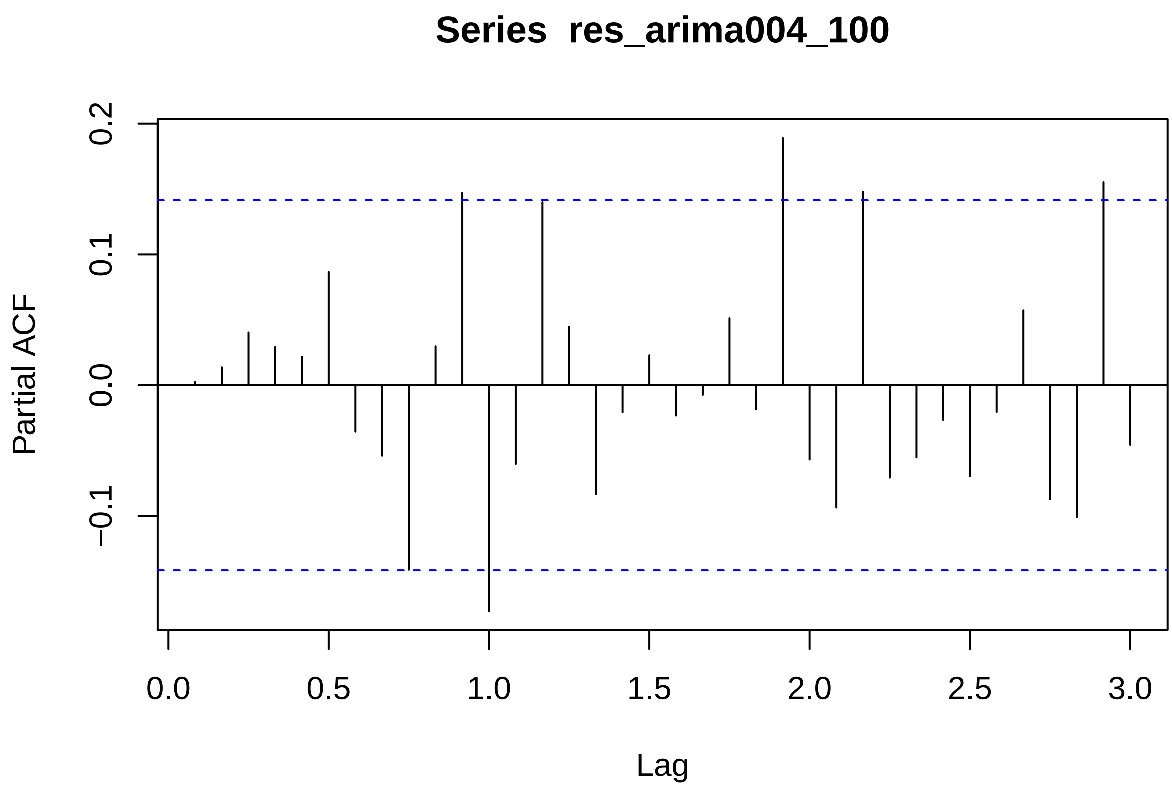
$ARIMA(0, 0, 4)(1, 0, 0)[12]$

```r
qns_arima004_100_fit_cons <- qns_ts3 %>%
  model(arima004_100_constant = ARIMA(total_waste ~ 1 +
                                    pdq(0,0,4) +
                                    PDQ(1,0,0, period = 12)))

zoo_arima004_100_fit <- arima(DSNY_QNS_zoo_ts,
                              order = 1 + c(0,0,4),
```

```
                              seasonal = list(order = c(1, 0L, 0L),
                                              period = 12))

res_arima004_100 <- zoo_arima004_100_fit$residuals
acf(res_arima004_100, lag.max = 36)
```




```
pacf(res_arima004_100, lag.max = 36)
```




```
accuracy(qns_arima004_100_fit_cons)[4]
```

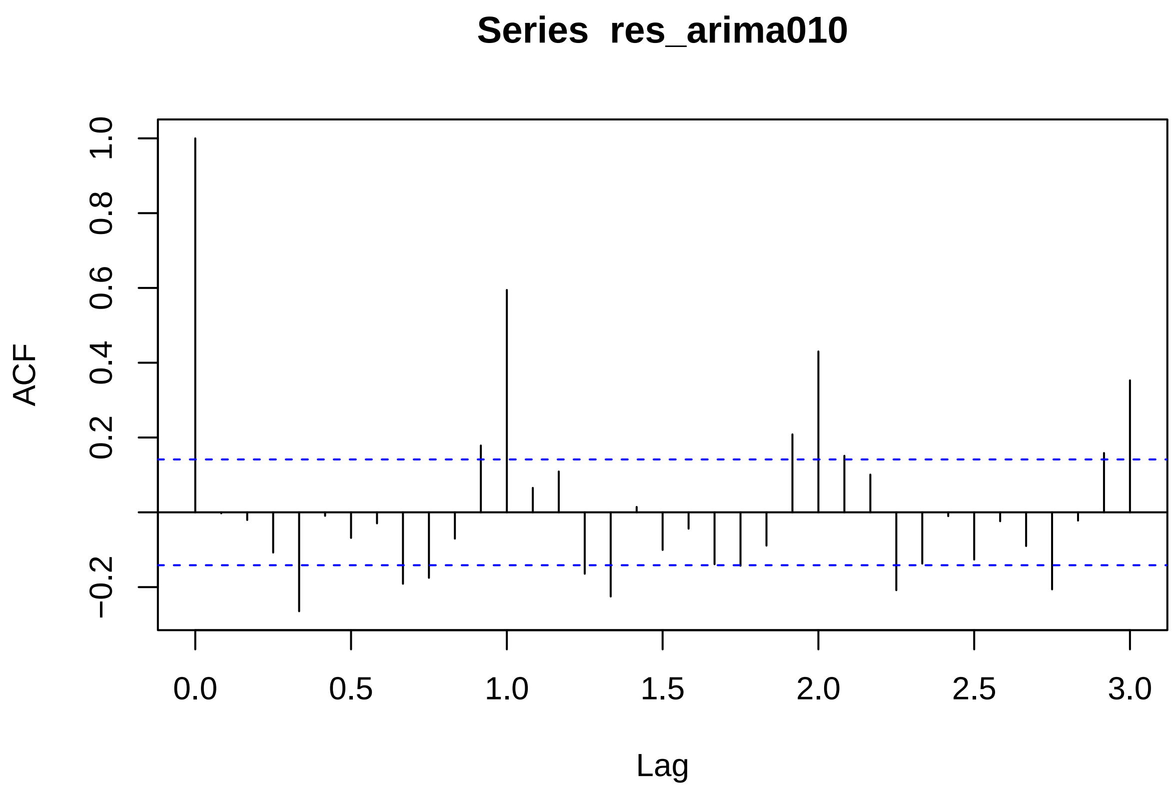
```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 4163.
```

RMSE = 4162.64. The first 10 lags are not significant in the PACF plot. The seasonal lags are negatively autocorrelated. Other lags are also significant. This model does not appear to produce a stable model, according to their plots.

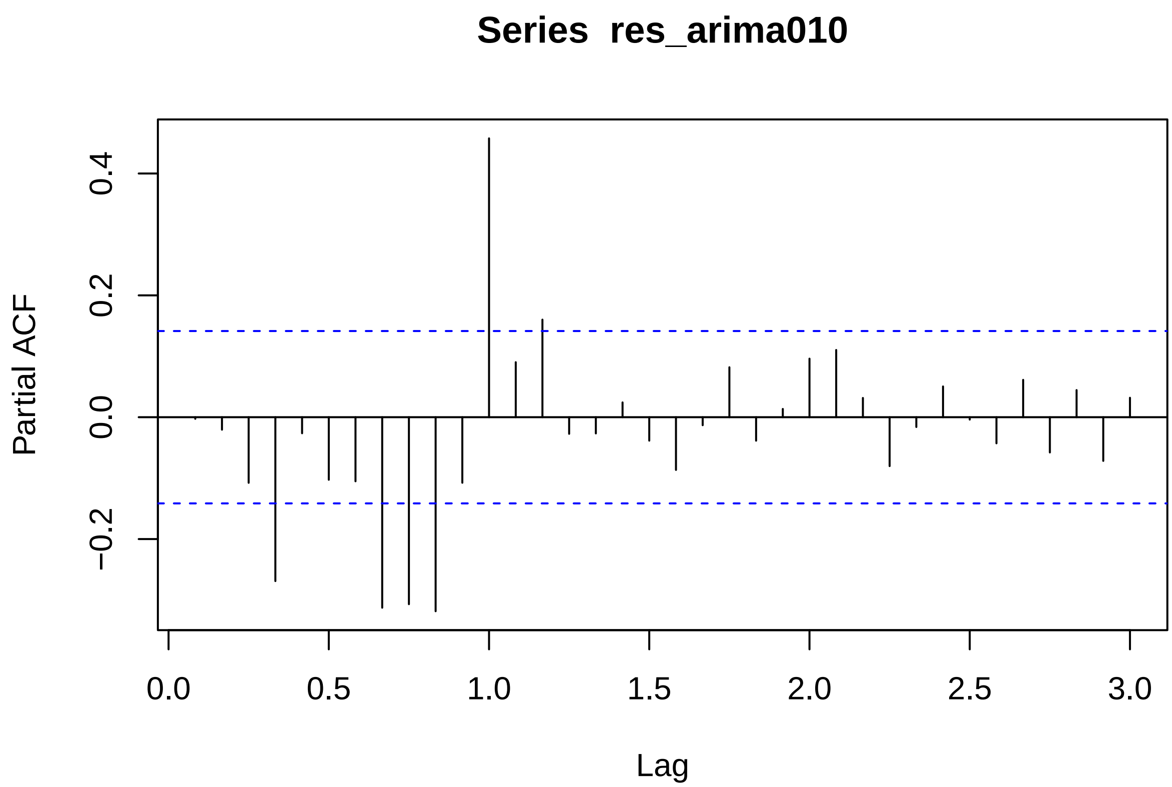
ARIMA(0,1,0) with constant

$ARIMA(0,1,0)$

```
qns_arima010_fit_cons <- qns_ts3 %>%
  model(arima010_constant = ARIMA(total_waste ~ 1 + pdq(0,1,0)))

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




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




```r
accuracy(qns_arima010_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 6502.
```

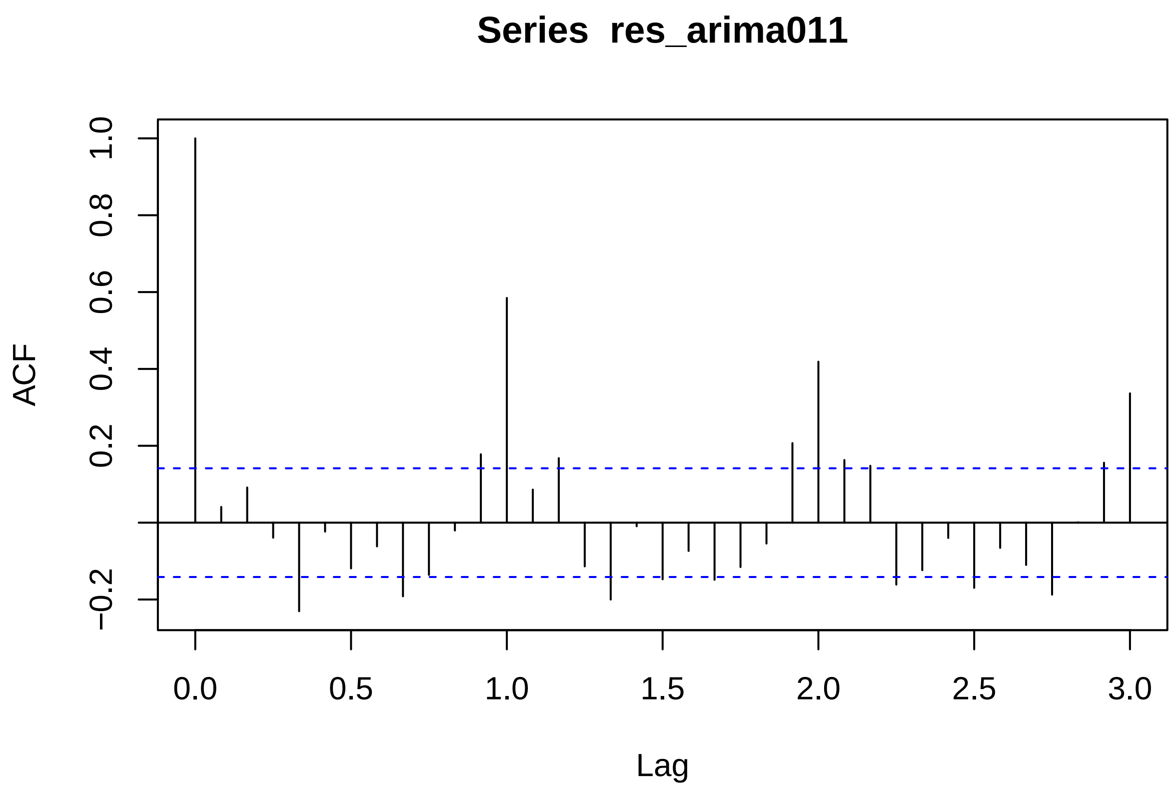
RMSE = 6502.128. In both the ACF and PACF plot, we do see a tiny lag-1 negative autocorrelation. According to rule 7, "The lag at which the ACF cuts off is the indicated number of MA terms." However, we also see that the first significant lag is lag-4, which is also negative. Lag-12 is positively correlated and significant.

Let's explore a MA(1) model

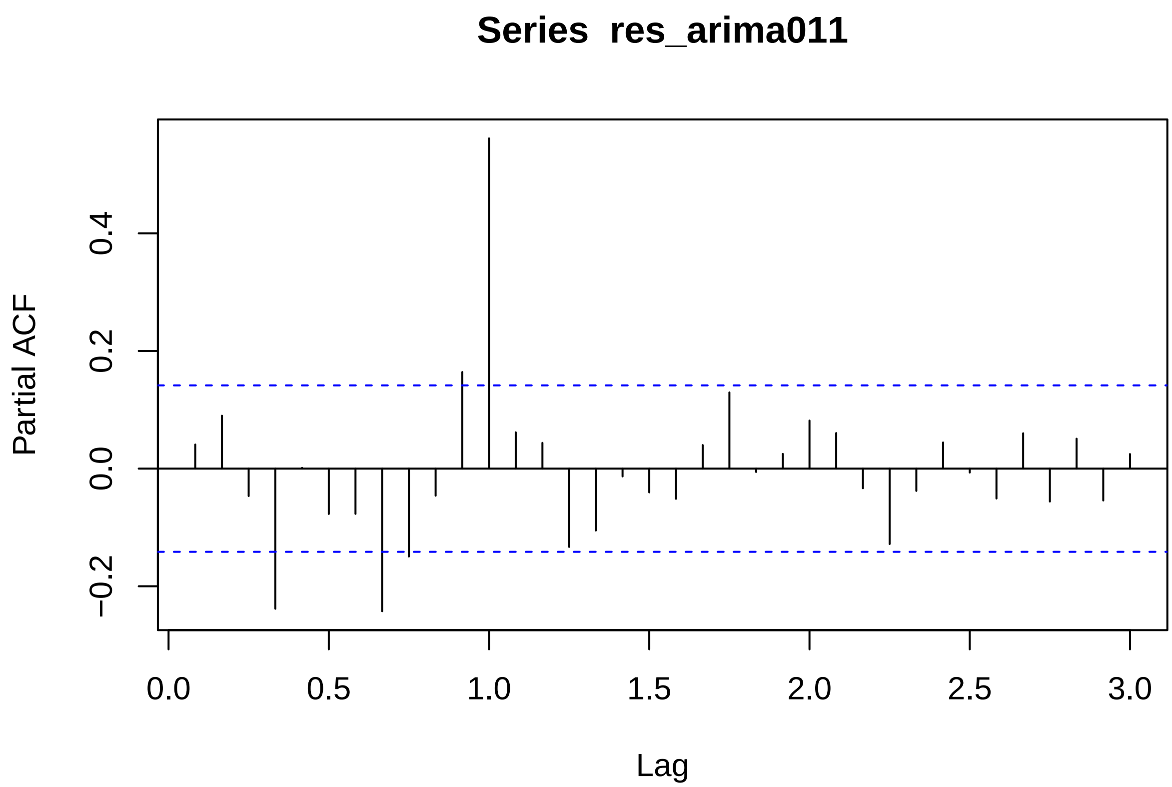
ARIMA(0,1,1) with constant

$ARIMA(0,1,1)$

```r
qns_arima011_fit_cons <- qns_ts3 %>%
  model(arima011_constant = ARIMA(total_waste ~ 1 + pdq(0,1,1)))

zoo_arima011_fit <- arima(DSNY_QNS_zoo_ts, order = 1 + c(0,1,1))
res_arima011 <- zoo_arima011_fit$residuals
acf(res_arima011, lag.max = 36)
```




```r
pacf(res_arima011, lag.max = 36)
```




```
accuracy(qns_arima011_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 6421.
```

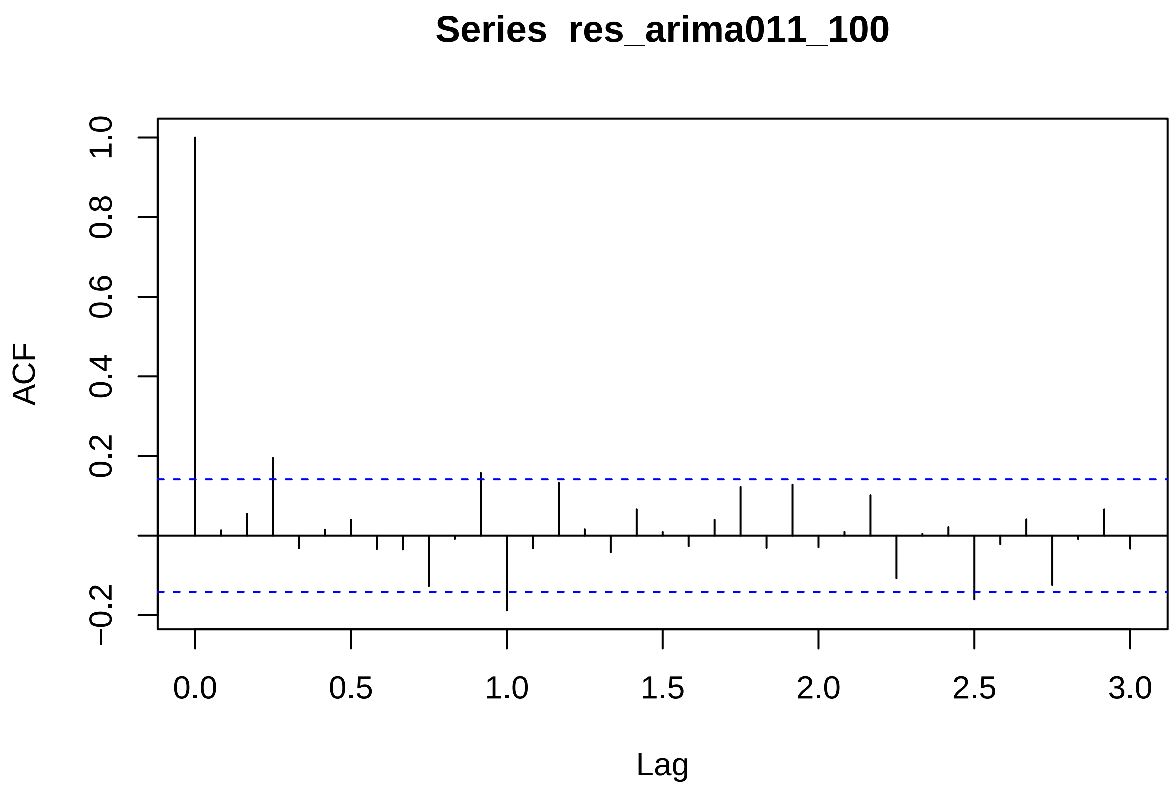
RMSE = 6421.289, which is a decrease from the base differenced model. The lags = (1,2,3) are still not significant and we see that lag=1 in the PACF has become positive. What we need to keep in mind is that the seasonal lags, lag = 12 is approximately 0.5. Its possible to add a seasonal difference, or add a seasonal AR(1).

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

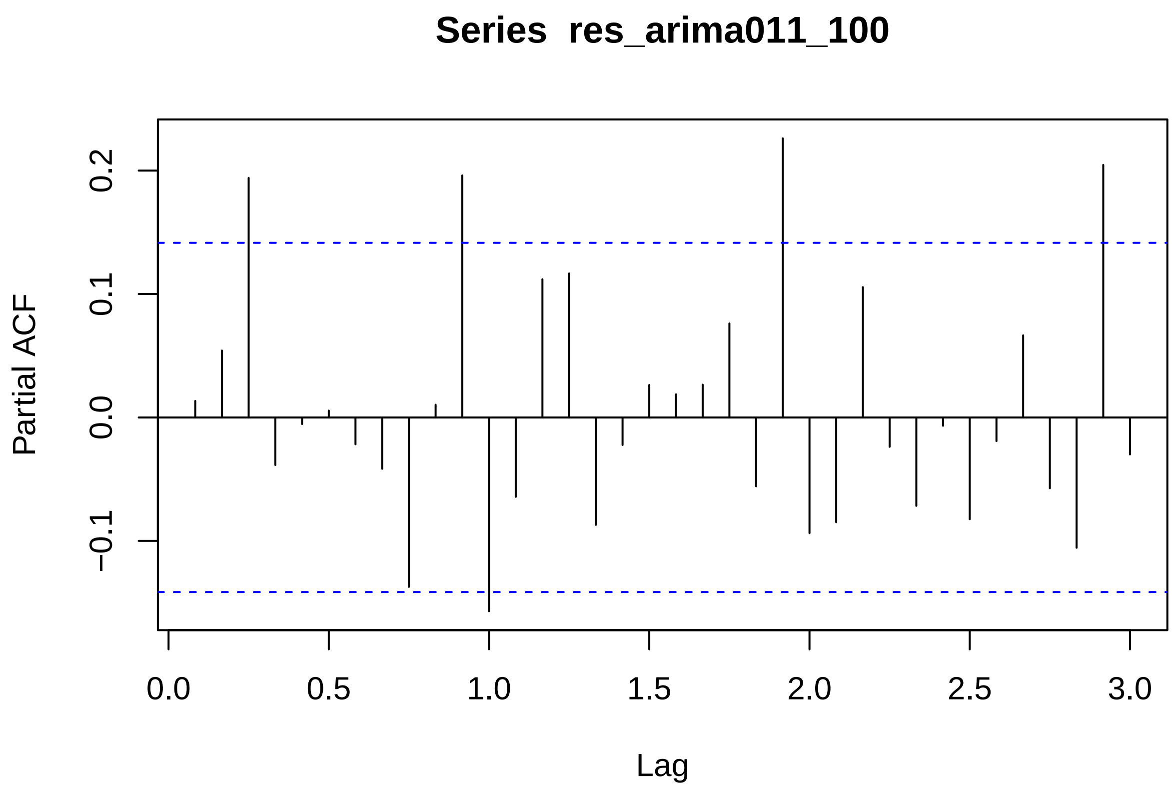
$ARIMA(0,1,1)(1,0,0)[12]$

```
qns_arima011_100_fit_cons <- qns_ts3 %>%
  model(arima011_100_constant = ARIMA(total_waste ~ 1 +
                                       pdq(0,1,1) +
                                       PDQ(1,0,0,
                                           period = 12)))

zoo_arima011_100_fit <- arima(DSNY_QNS_zoo_ts,
                              order = 1 + c(0,1,1),
```

```
                              seasonal = list(order = c(1,0L,0L),
                                              period = 12))
res_arima011_100 <- zoo_arima011_100_fit$residuals
acf(res_arima011_100, lag.max = 36)
```




```
pacf(res_arima011_100, lag.max = 36)
```




```
accuracy(qns_arima011_100_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 4174.
```

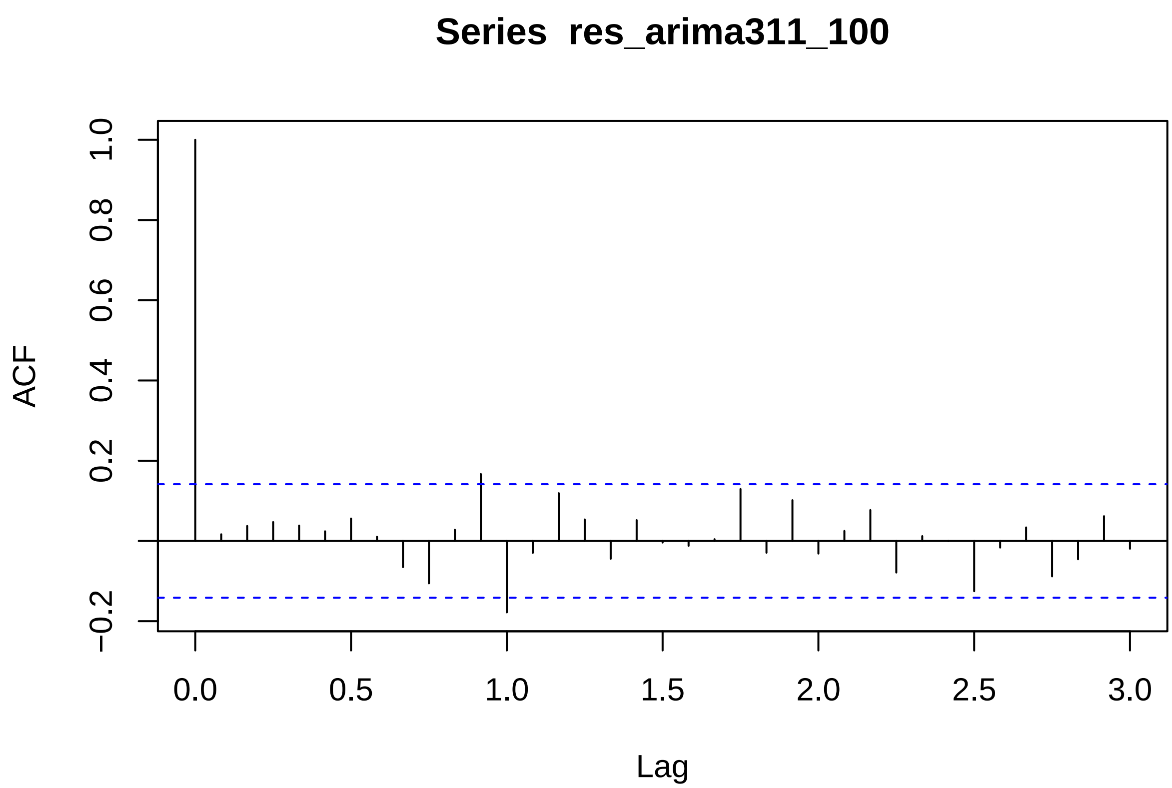
A decrease in the RMSE = 4174.407. Now lag = 3 is significant and positively correlated. Lags = 11, 23, 35 are positive and significant. We also see that lag = 12, the seasonal lag is also significant. Although we see a decrease in the RMSE, the plots of PACF show mixing autocorrelations at high order lags.

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

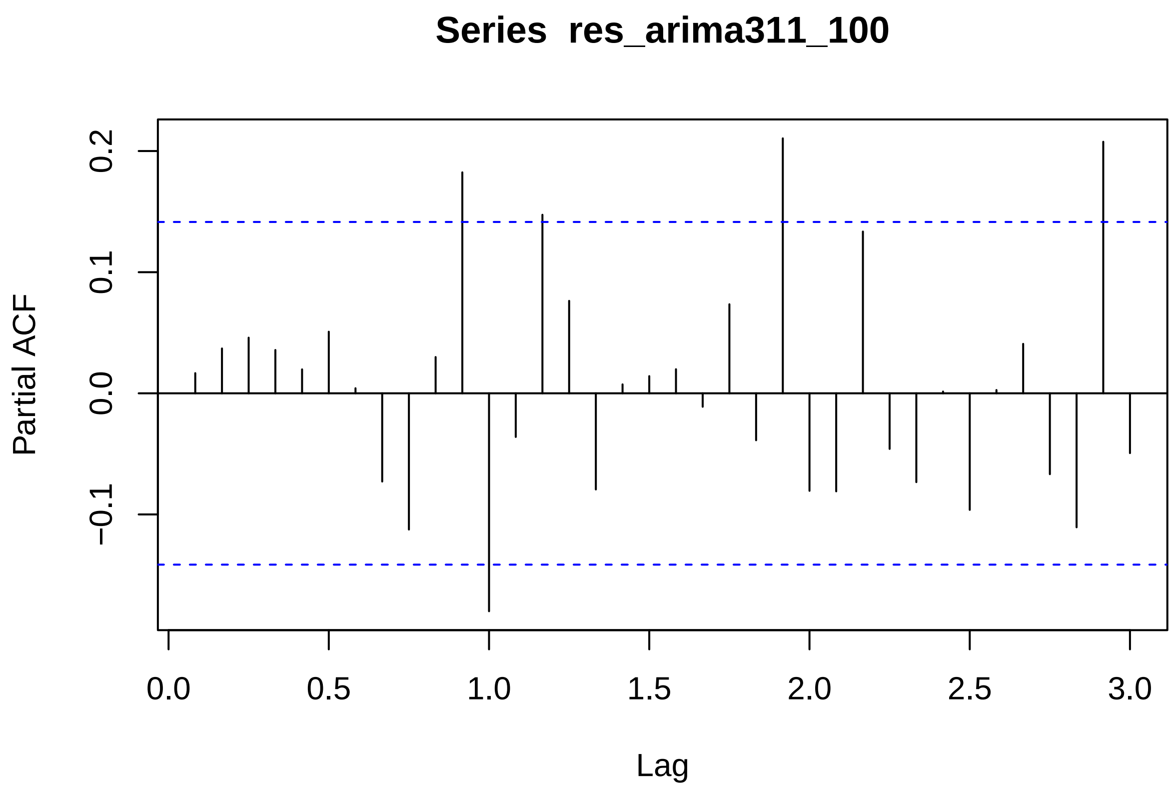
$ARIMA(3,1,1)(1,0,0)[12]$

```
qns_arima311_100_fit_cons <- qns_ts3 %>%
  model(arima311_100_constant = ARIMA(total_waste ~ 1 +
                                        pdq(3,1,1) +
                                        PDQ(1,0,0,
                                            period = 12)))

zoo_arima311_100_fit <- arima(DSNY_QNS_zoo_ts,
                              order = 1 + c(3,1,1),
                              seasonal = list(order = c(1,0L,0L),
```

```
                                              period = 12))
res_arima311_100 <- zoo_arima311_100_fit$residuals
acf(res_arima311_100, lag.max = 36)
```




```
pacf(res_arima311_100, lag.max = 36)
```



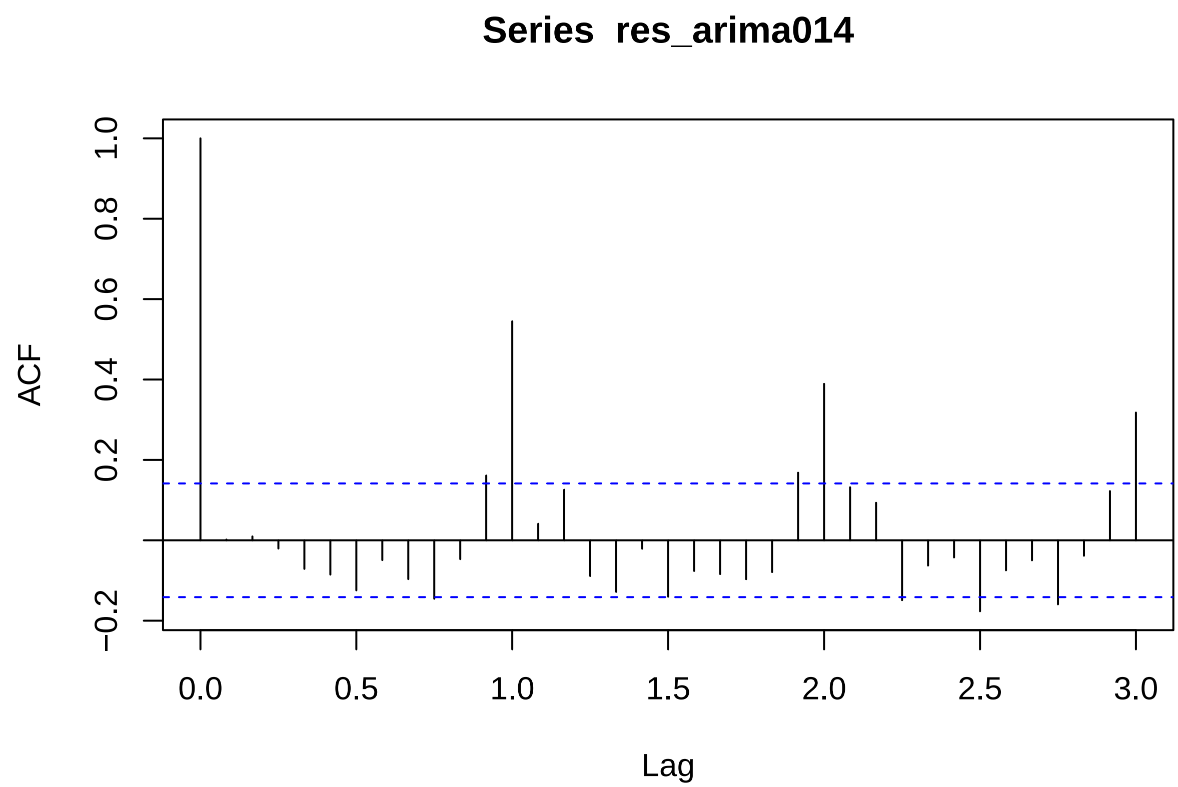

```
accuracy(qns_arima311_100_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 4112.
```

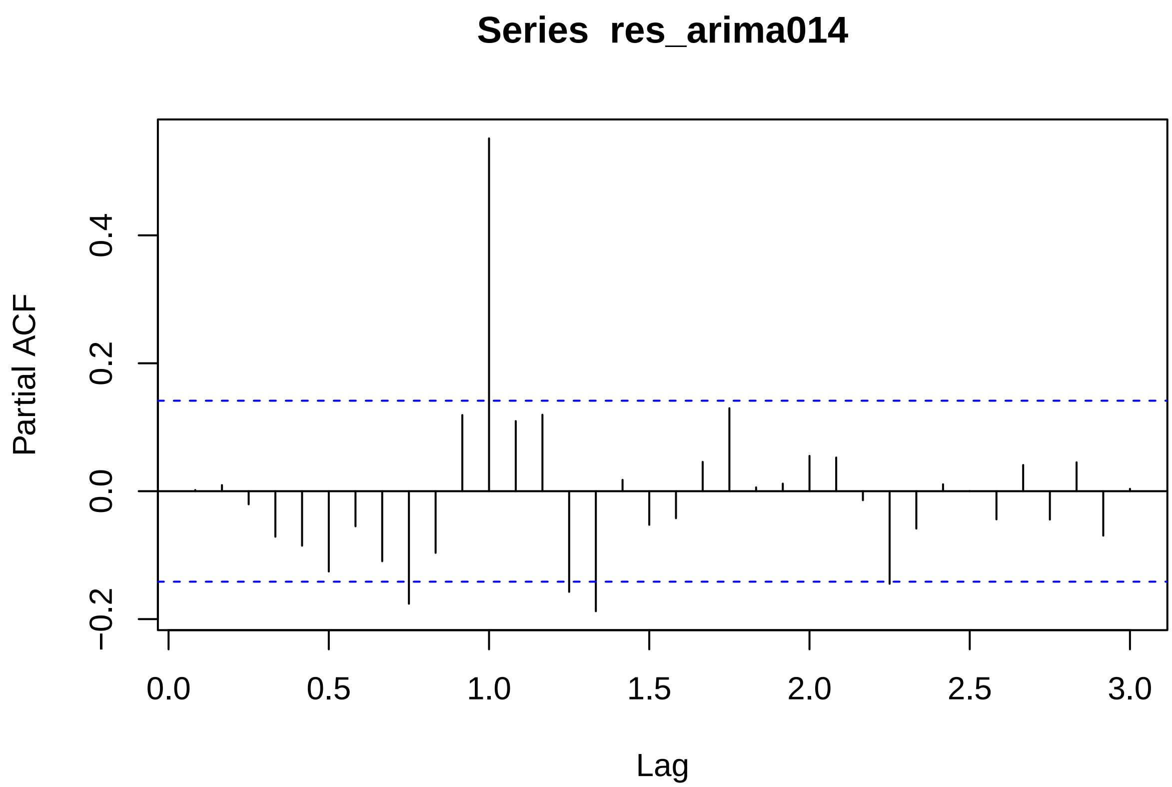
ARIMA(0,1,4) with constant

$ARIMA(0,1,4)$

```
qns_arima014_fit_cons <- qns_ts3 %>%
  model(arima014_constant = ARIMA(total_waste ~ 1 + pdq(0,1,4)))

zoo_arima014_fit <- arima(DSNY_QNS_zoo_ts, order = 1 + c(0,1,4))
res_arima014 <- zoo_arima014_fit$residuals
acf(res_arima014, lag.max = 36)
```




```r
pacf(res_arima014, lag.max = 36)
```




```
accuracy(qns_arima014_fit_cons)[4]
```

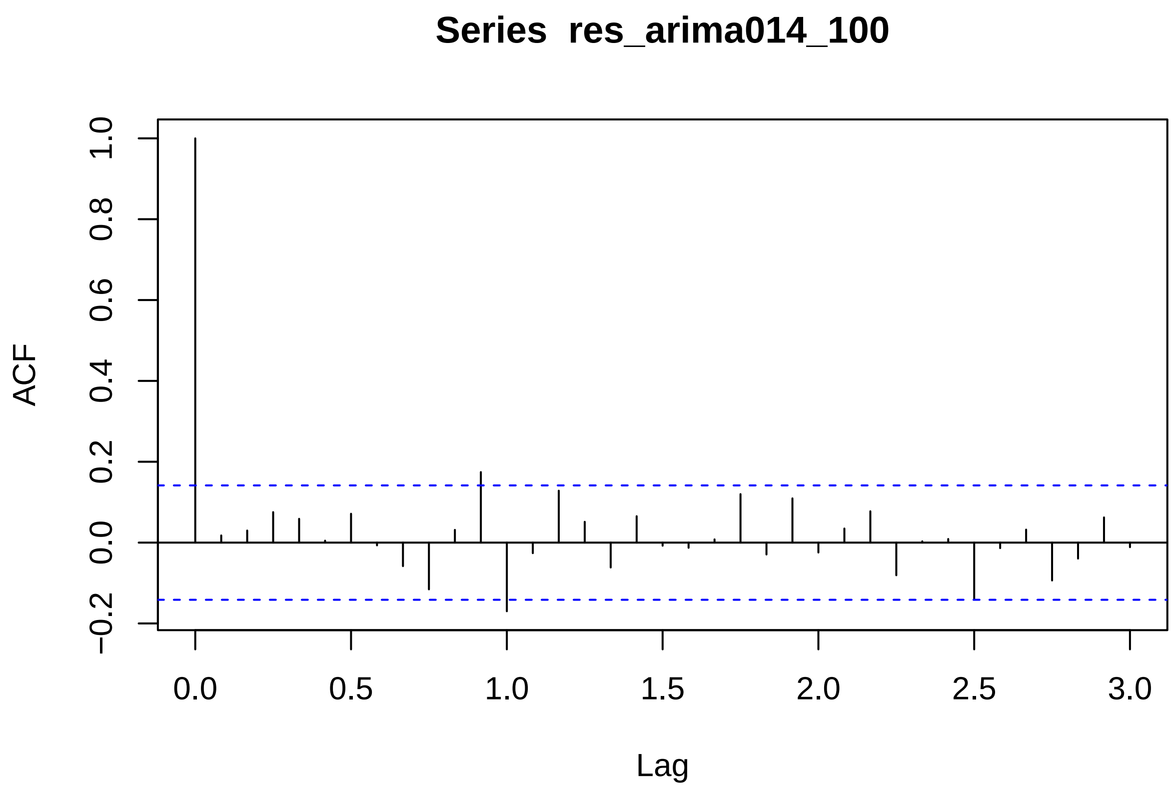
```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 5649.
```

RMSE = 5649.113

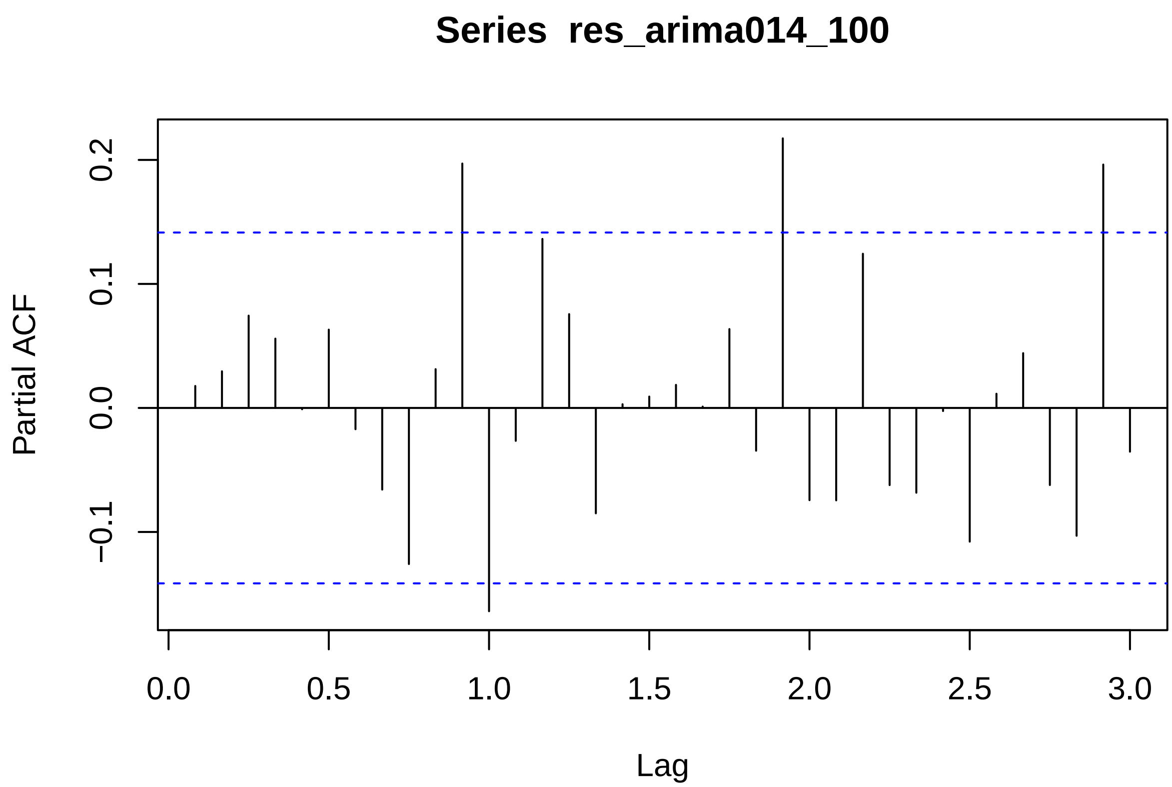
ARIMA(0,1,4)(1,0,0)[12] with constant and seasonality

$ARIMA(0,1,4)(1,0,0)[12]$

```
qns_arima014_100_fit_cons <- qns_ts3 %>%
  model(arima014_100_constant = ARIMA(total_waste ~ 1 +
                                        pdq(0,1,4) +
                                        PDQ(1,0,0,
                                            period = 12)))

zoo_arima014_100_fit <- arima(DSNY_QNS_zoo_ts,
                              order = 1 + c(0,1,4),
                              seasonal = list(order = c(1,0L,0L),
                                              period = 12))
res_arima014_100 <- zoo_arima014_100_fit$residuals
acf(res_arima014_100, lag.max = 36)
```




```
pacf(res_arima014_100, lag.max = 36)
```




```
accuracy(qns_arima014_100_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 4123.
```

RMSE = 4123.31. Bounded b/w (-0.2 and 0.2)

AutoArima

```
qns_auto_arima_fit_cons <- qns_ts3 %>%
  model(stepwise = ARIMA(total_waste),
        search = ARIMA(total_waste,
                       stepwise = FALSE,
                       approximation = FALSE))
accuracy(qns_auto_arima_fit_cons)[1:4]
```

```
## # A tibble: 2 x 4
##   .model   .type       ME  RMSE
##   <chr>    <chr>    <dbl> <dbl>
## 1 stepwise Training  11.2 5494.
## 2 search   Training  18.2 5477.
```

```r
qns_auto_arima_fit_cons %>% select(.model = stepwise) %>% report()
```

```
## Series: total_waste
## Model: ARIMA(2,0,3) w/ mean
##
## Coefficients:
##          ar1      ar2      ma1     ma2     ma3    constant
##       0.7094  -0.6198  -0.1148  0.6793  0.3247  67042.0634
## s.e.  0.1187   0.1007   0.1314  0.0536  0.0895    746.8235
##
## sigma^2 estimated as 31157027:  log likelihood=-1926.61
## AIC=3867.22   AICc=3867.83   BIC=3890.02
```

```r
print("--------------")
```

```
## [1] "--------------"
```

```r
qns_auto_arima_fit_cons %>% select(.model = search) %>% report()
```

```
## Series: total_waste
## Model: ARIMA(3,0,2) w/ mean
##
## Coefficients:
##          ar1      ar2     ar3     ma1     ma2    constant
##       1.1361  -1.0305  0.3128  -0.539  0.8876  42828.2096
## s.e.  0.0982   0.1151  0.0828   0.059  0.0485    530.5093
##
## sigma^2 estimated as 30965765:  log likelihood=-1926.01
## AIC=3866.03   AICc=3866.63   BIC=3888.83
```

Summary of models

$ARIMA(0,0,0)(1,0,0)[12]$ has $RMSE \approx 4413.677$
$ARIMA(3,0,0)(1,0,0)[12]$ has $RMSE \approx 4115.02$
$ARIMA(0,0,4)(1,0,0)[12]$ has $RMSE \approx 4162.64$
$ARIMA(3,1,1)(1,0,0)[12]$ has $RMSE \approx 4111.87$
$ARIMA(3,0,2)$ has $RMSE \approx 5477.051$