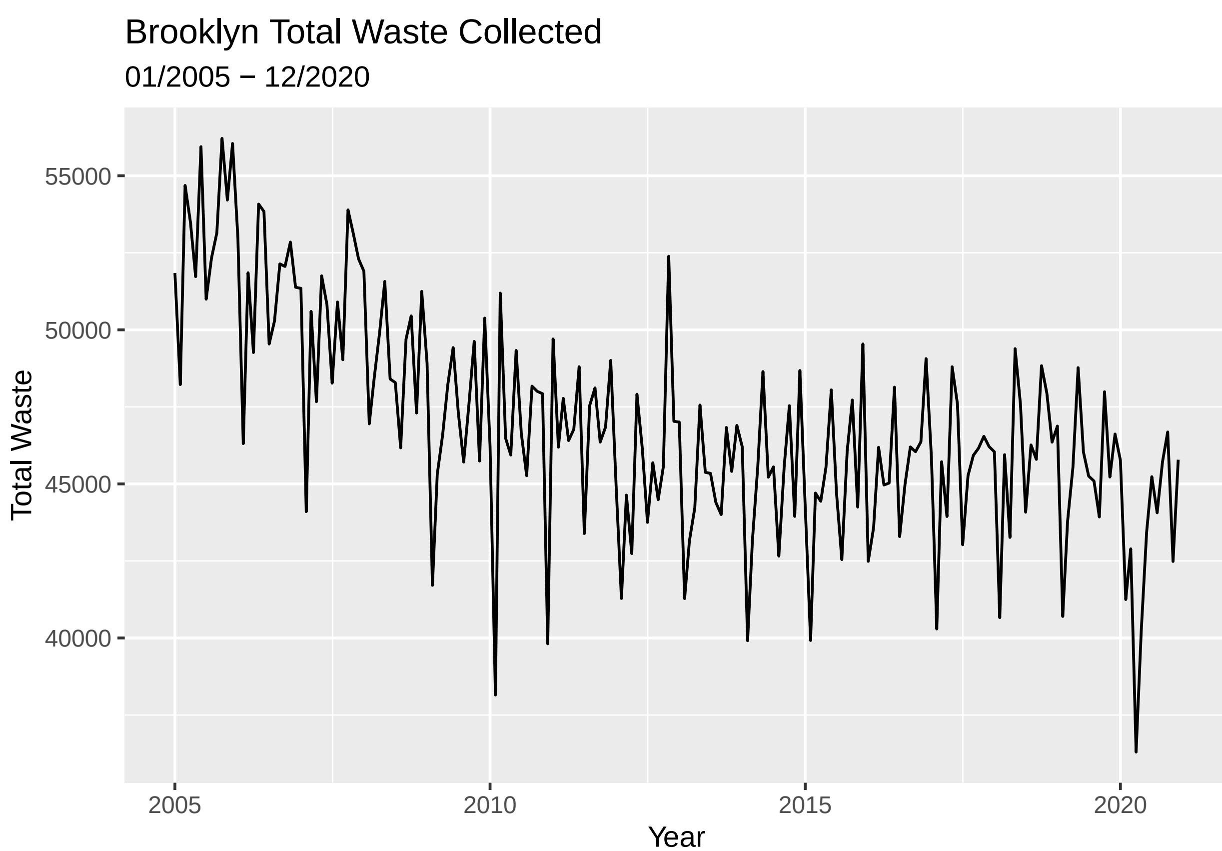
MN_ts.Rmd

Daniel L.

5/3/2022

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



I should investigate the average tonnage collected in MN from 2005 through 2010.

```
DSNY_third_manhattan %>%
  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                53237.
##  2  2006                51380.
##  3  2007                50315.
##  4  2008                49194.
##  5  2009                47211.
##  6  2010                46094.
##  7  2011                47242.
##  8  2012                45549.
##  9  2013                45123.
## 10  2014                45221.
## 11  2015                45143.
## 12  2016                45528.
## 13  2017                45445.
## 14  2018                46012.
## 15  2019                45485.
## 16  2020                43321.
```

We do see about a decrease in average waste collected by 7,000 tons in 2010, when compared to 2005. And a decrease in average waste collected by 4,000 tons in 2020, when compared to 2011.

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

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1      2.66        0.01
```

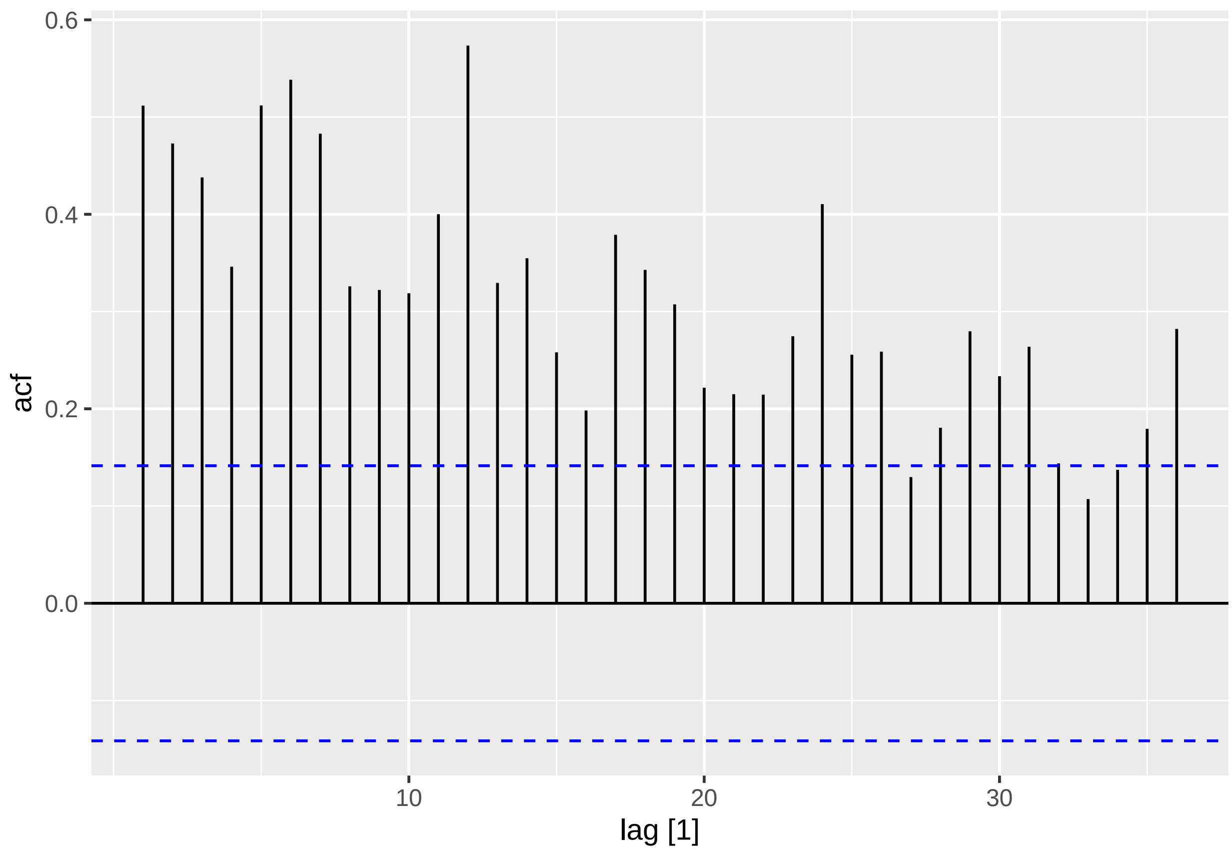
```
#differenced values
man_ts %>% features(diff1, unitroot_kpss)
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1    0.0259         0.1
```

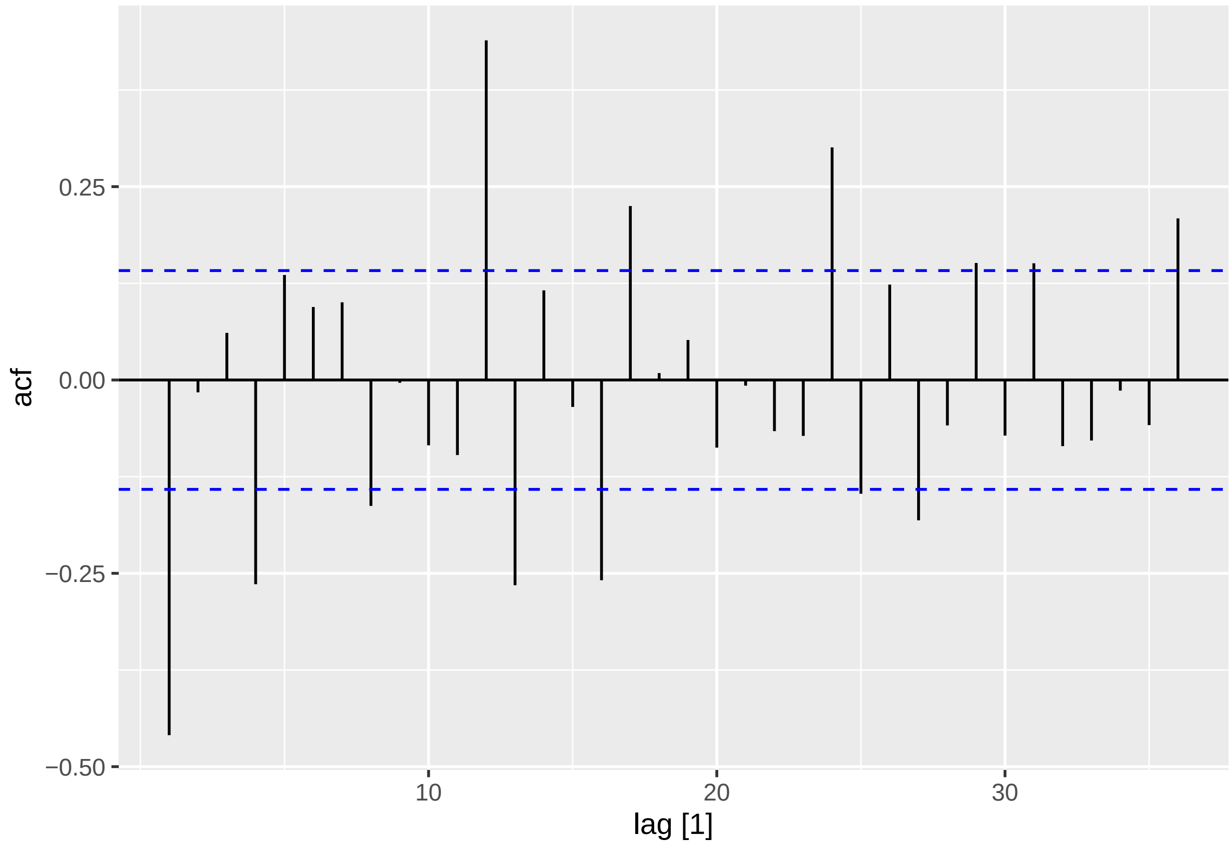
According to the results of the KPSS test, we reject the H_0 when evaluating the total_waste values. We fail to reject the H_0 when evaluating the differenced values

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

```
man_ts3 %>%
  ACF(total_waste, lag_max = 36) %>%
  autoplot()
```

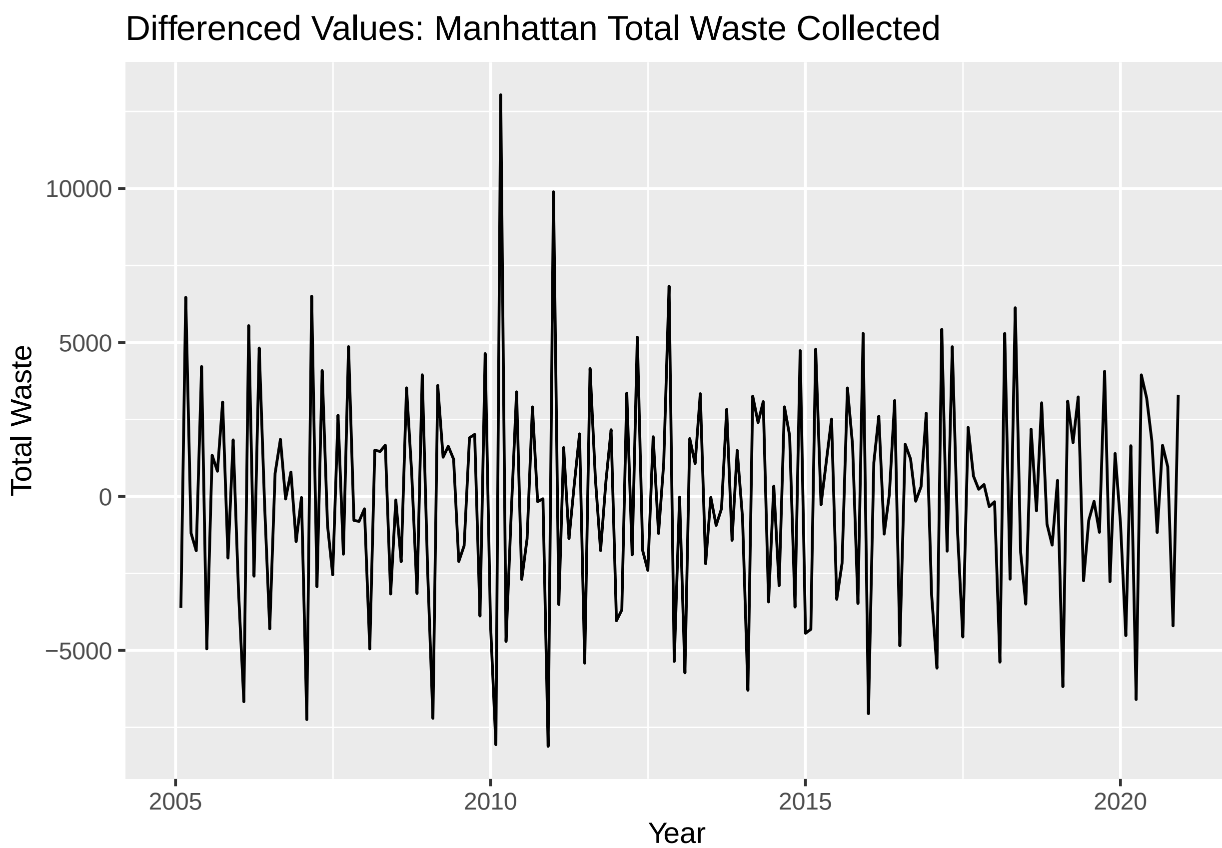



```
#acf of the differenced values
man_ts3 %>%
  ACF(diff1, lag_max = 36) %>%
  autoplot()
```




```
man_ts3 %>%
  ggplot(mapping = aes(x = month, y = diff1)) + geom_line() +
  labs(x = "Year",
       y = "Total Waste",
       title = "Differenced Values: Manhattan Total Waste Collected")
```

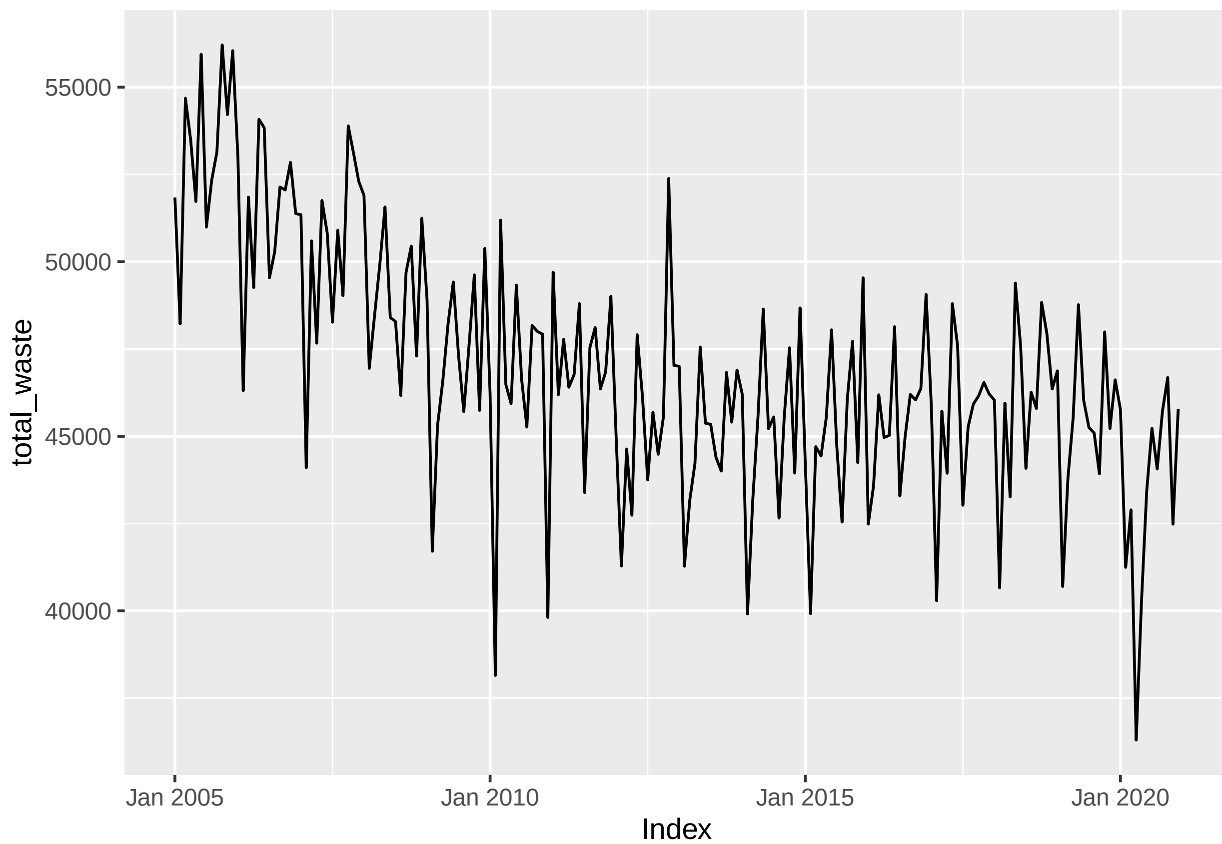
```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



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

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

autoplot(as.zoo(DSNY_MN_zoo_ts))
```

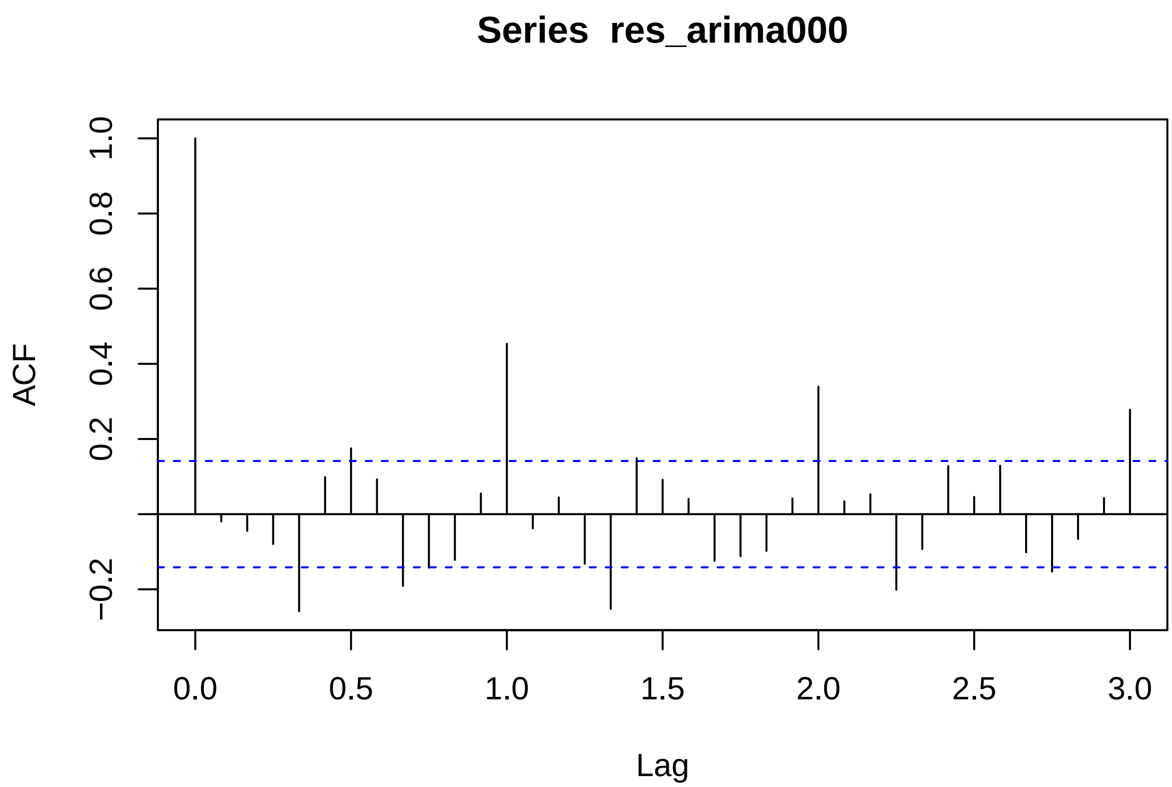


ARIMA(0,0,0) with constant

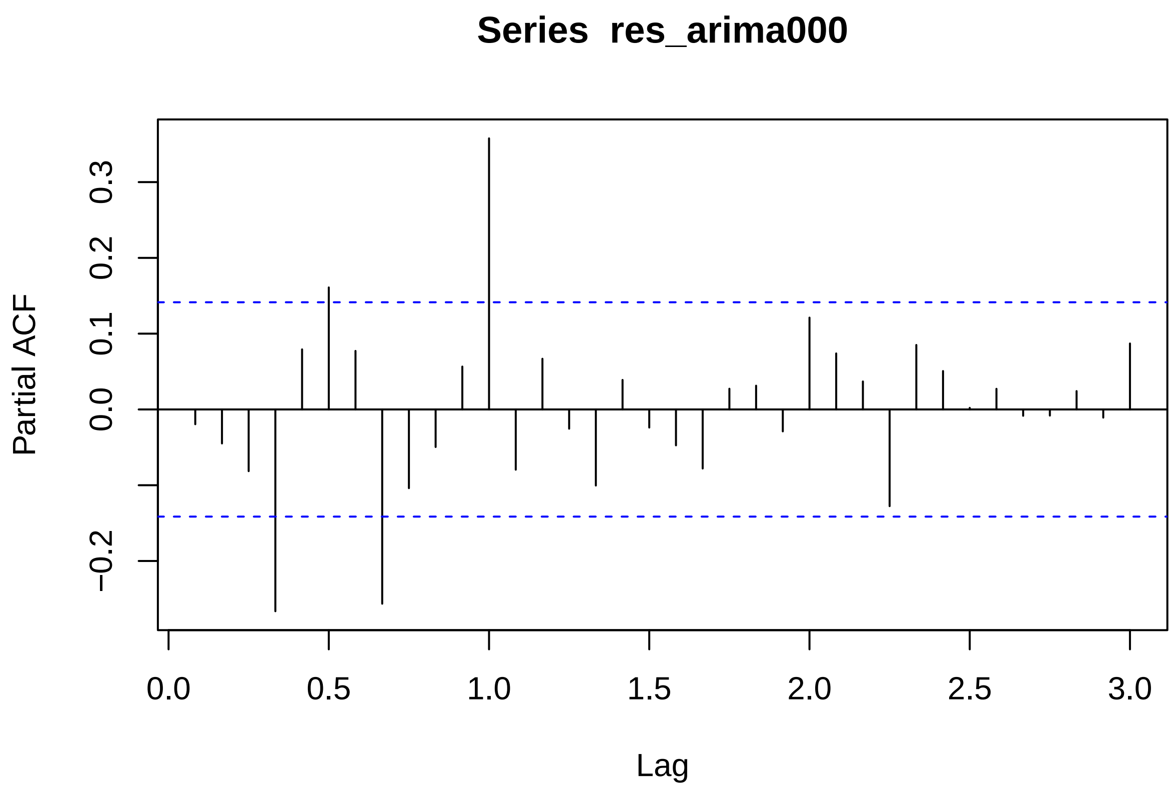
$ARIMA(0,0,0)$

```
man_arima000_fit_cons <- man_ts3 %>%
  model(arima000_constant = ARIMA(total_waste ~ 1 + pdq(0,0,0)))

zoo_arima000_fit <- arima(DSNY_MN_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(man_arima000_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 3502.
```

RMSE = 3501. The first significant lag in the ACF plot is lag 4. The first significant lag in the PACF plot is also lag 4. Both these lags are negative, which indicate the use of an MA() argument. The seasonal lags are once again present in both plots. In the PACF() plot, lag 6 is positive and significant.

With out working with the differenced values yet, I will add the MA() or seasonal MA paramters first.

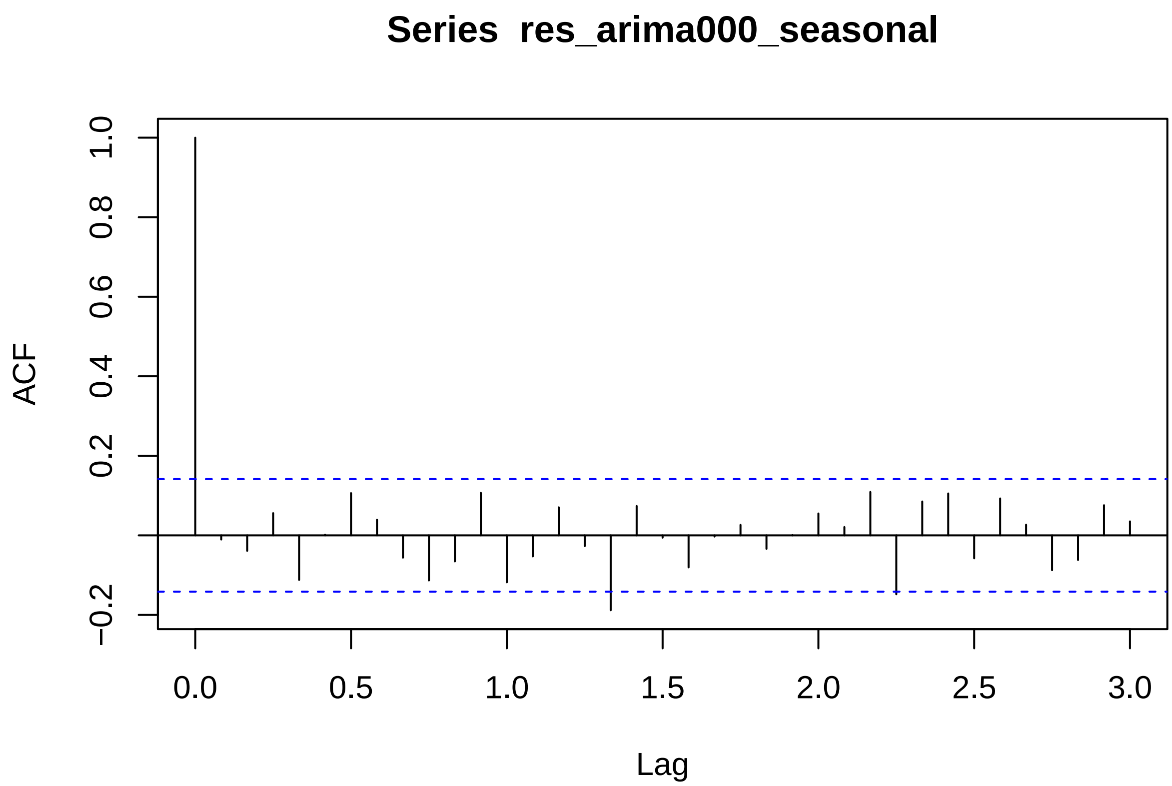
ARIMA(0,0,0)(1,0,0) with constant and seasonal parameter

$ARIMA(0, 0, 0)(1, 0, 0)$

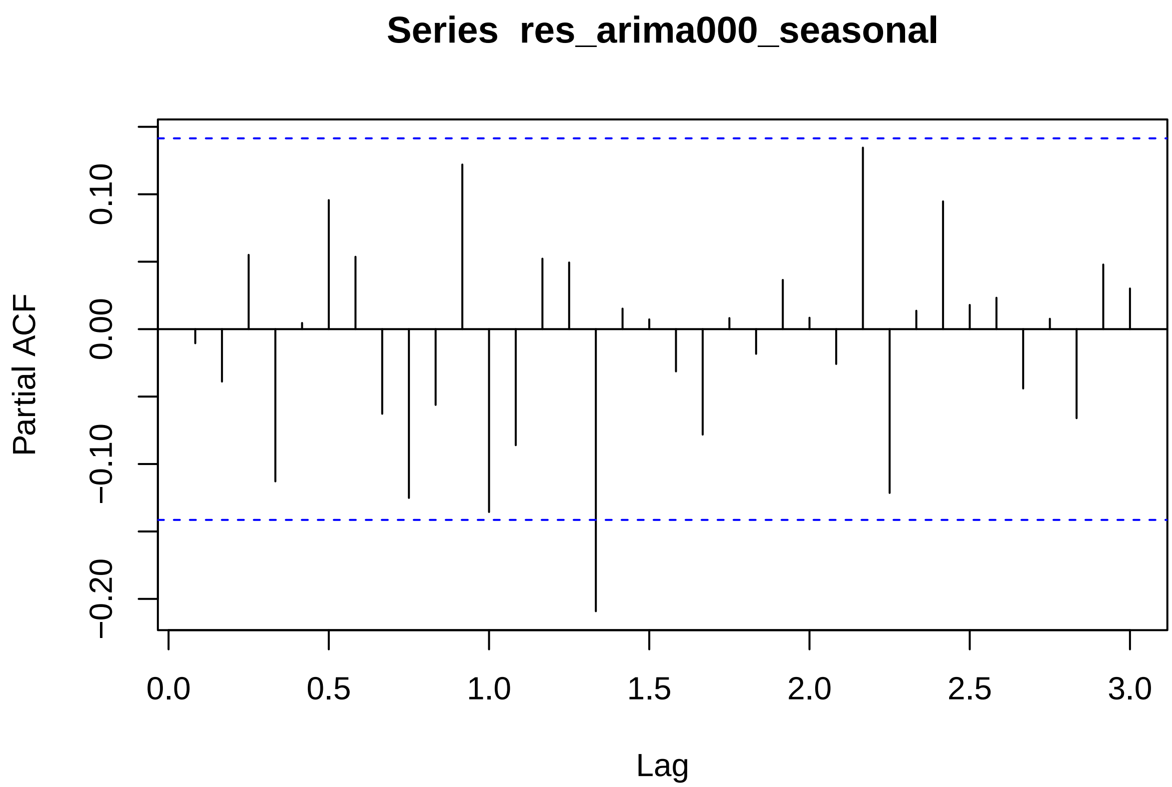
```
man_arima000_fit_seasonal_cons <- man_ts3 %>%
  model(arima000_constant_seasonal = ARIMA(total_waste ~ 1 +
                                    pdq(0,0,0) +
                                    PDQ(1,0,0, period = 12)))

zoo_arima000_seasonal_fit <- arima(DSNY_MN_zoo_ts,
                          order = 1 + c(0,0,0),
```

```
                                seasonal = list(order = c(1,0L,0L), period = 12))
#names(zoo_arima000_fit)
res_arima000_seasonal <- zoo_arima000_seasonal_fit$residuals
acf(res_arima000_seasonal, lag.max = 36)
```




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




```
accuracy(man_arima000_fit_seasonal_cons)[4]
```

```
## # A tibble: 1 x 1
##     RMSE
##    <dbl>
## 1 2499.
```

RMSE = 2499.473. The majority of the lags in the ACF plot are contained within bounds. Along with the lags of the PACF plot. Only lag = 14 is significant and positive. The values are bounded b/w (-0.20, 0.15).

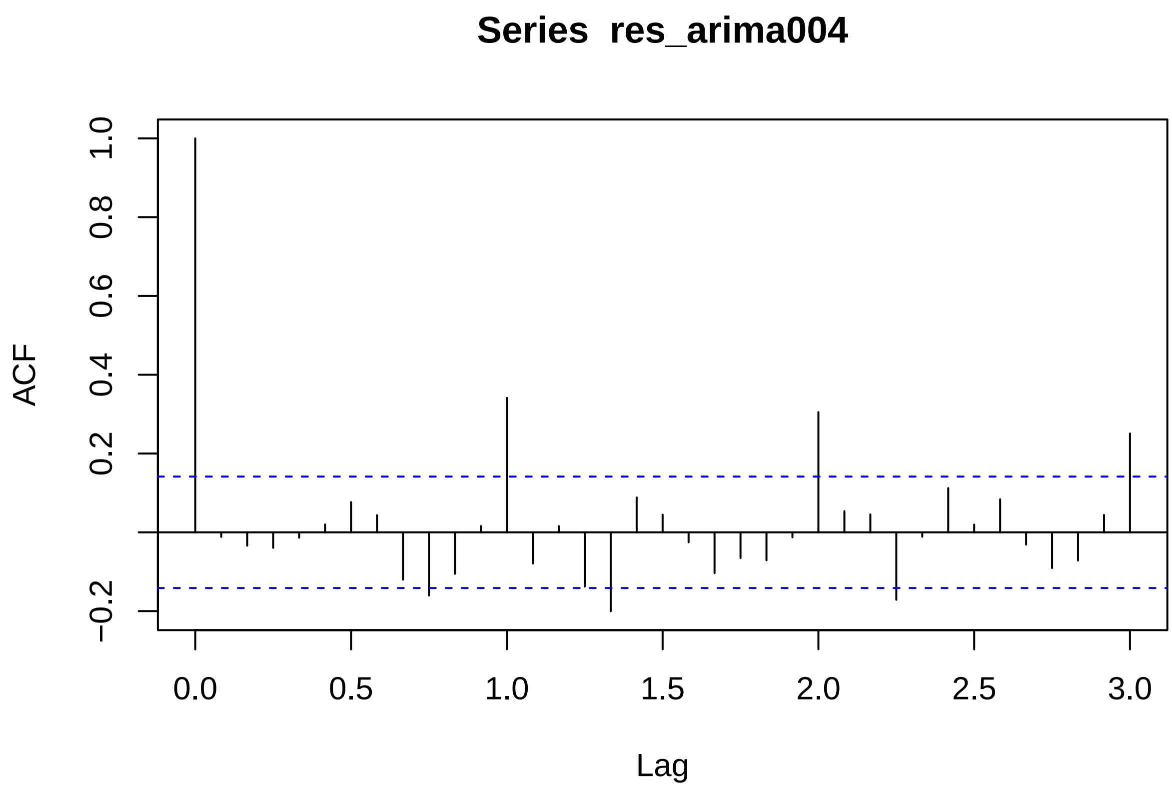
Let's work with an MA(4) model before we are confident in the previous model

ARIMA(0,0,4) with constant

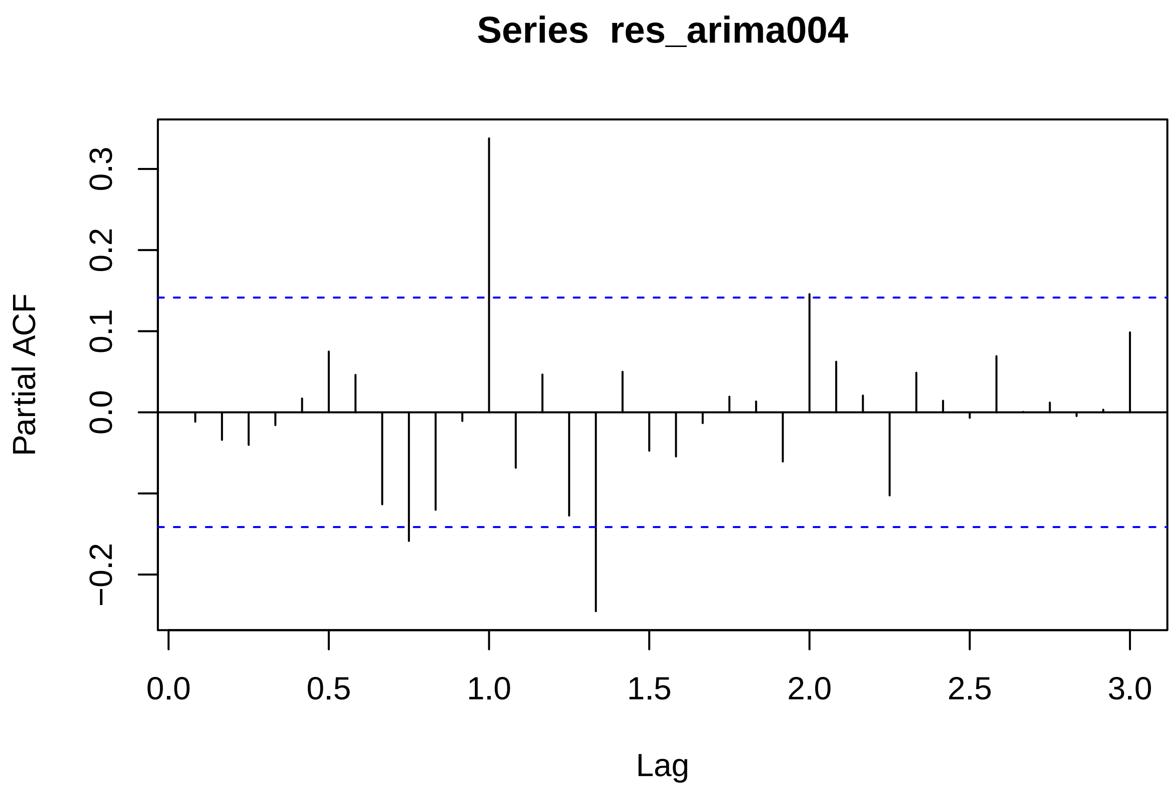
$ARIMA(0, 0, 4)$

```
man_arima004_fit_cons <- man_ts3 %>%
  model(arima004_constant = ARIMA(total_waste ~ 1 +
                                    pdq(0,0,4)))

zoo_arima004_fit <- arima(DSNY_MN_zoo_ts,
                          order = 1 + c(0,0,4))
#names(zoo_arima000_fit)
res_arima004 <- zoo_arima004_fit$residuals
acf(res_arima004, lag.max = 36)
```




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




```
accuracy(man_arima004_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 2862.
```

The RMSE = 2826.278. We do see a decrease in the RMSE when compared to the first ARIMA(0,0,0).The seasonal lags are significant in both plots. In the PACF plot, lags 1-12 are not significant and contained within the bounds.

Lets work with this model and add a seasonal argument.

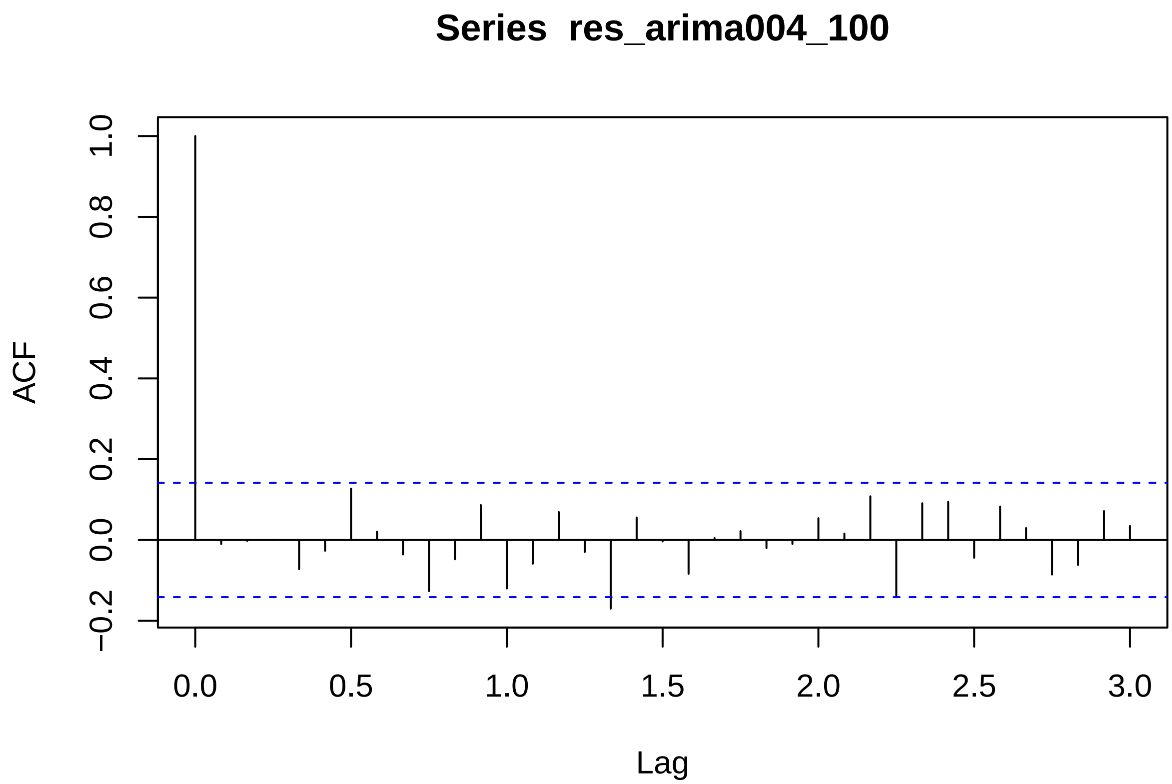
ARIMA(0,0,4)(1,0,0) with constant and seasonal

$ARIMA(0,0,4)(1,0,0)_{12}$

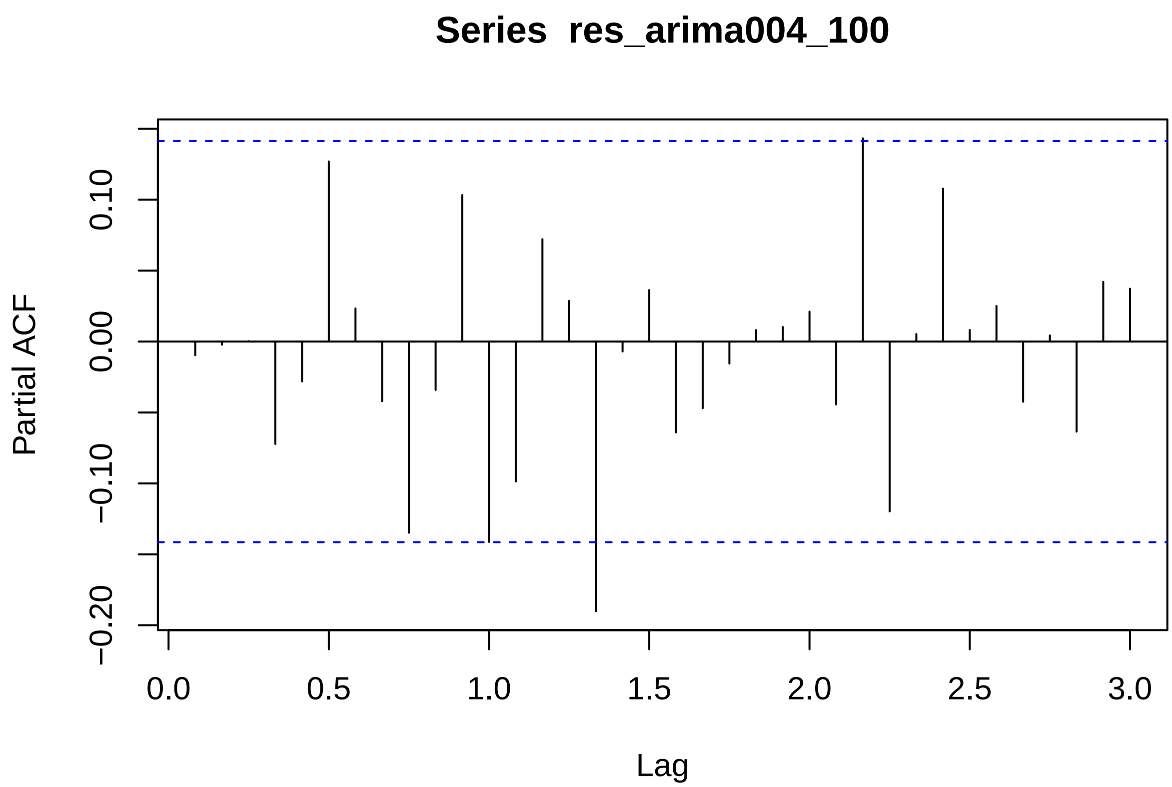
```
man_arima004_100_seasonal_fit_cons <- man_ts3 %>%
  model(arima004_constant = ARIMA(total_waste ~ 1 +
                                    pdq(0,0,4) +
                                    PDQ(1,0,0,
                                        period = 12)))

zoo_arima004_100_seasonalfit <- arima(DSNY_MN_zoo_ts,
```

```r
                           order = 1 + c(0,0,4),
                           seasonal = list(order = c(1, 0L, 0L), period = 12))
#names(zoo_arima000_fit)
res_arima004_100 <- zoo_arima004_100_seasonalfit$residuals
acf(res_arima004_100, lag.max = 36)
```




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




```
accuracy(man_arima004_100_seasonal_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 2386.
```

RMSE = 2386.418. In both plots, all first 16 lags are not significant. It is hard to explain why lag = 12 in the PACF plot is significant. This model is a strong contender.

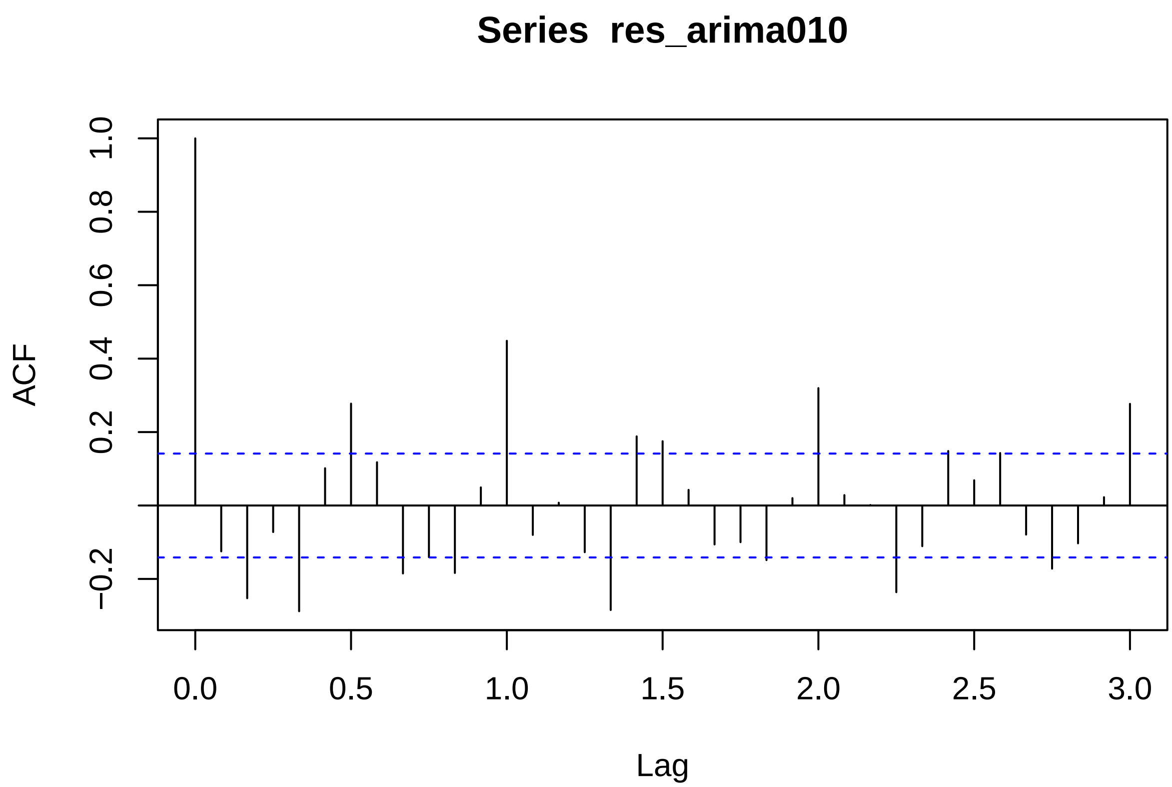
Since we can also work with the differenced values, we will create some models with them

ARIMA(0,1,0) with constant

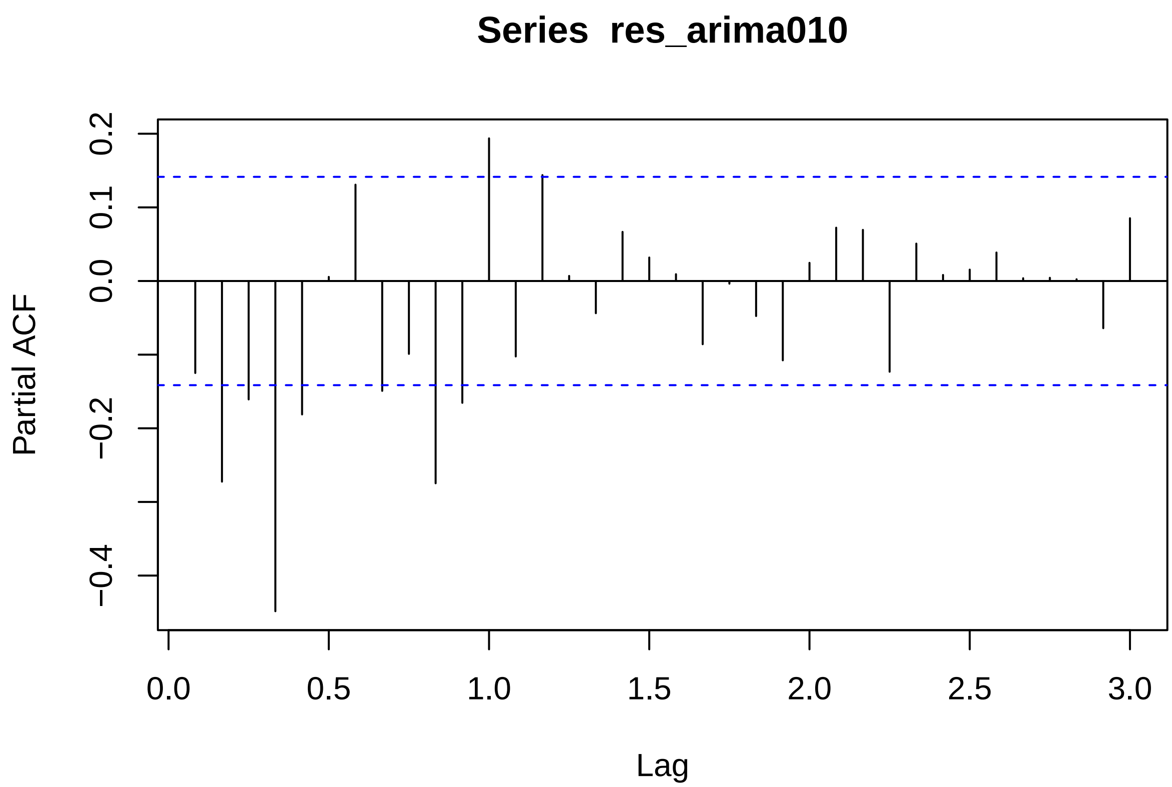
$ARIMA(0,1,0)$

```
man_arima010_fit_cons <- man_ts3 %>%
  model(arima010_constant = ARIMA(total_waste ~ 1 + pdq(0,1,0)))

zoo_arima010_fit <- arima(DSNY_MN_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)
```




```
accuracy(man_arima010_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 3441.
```

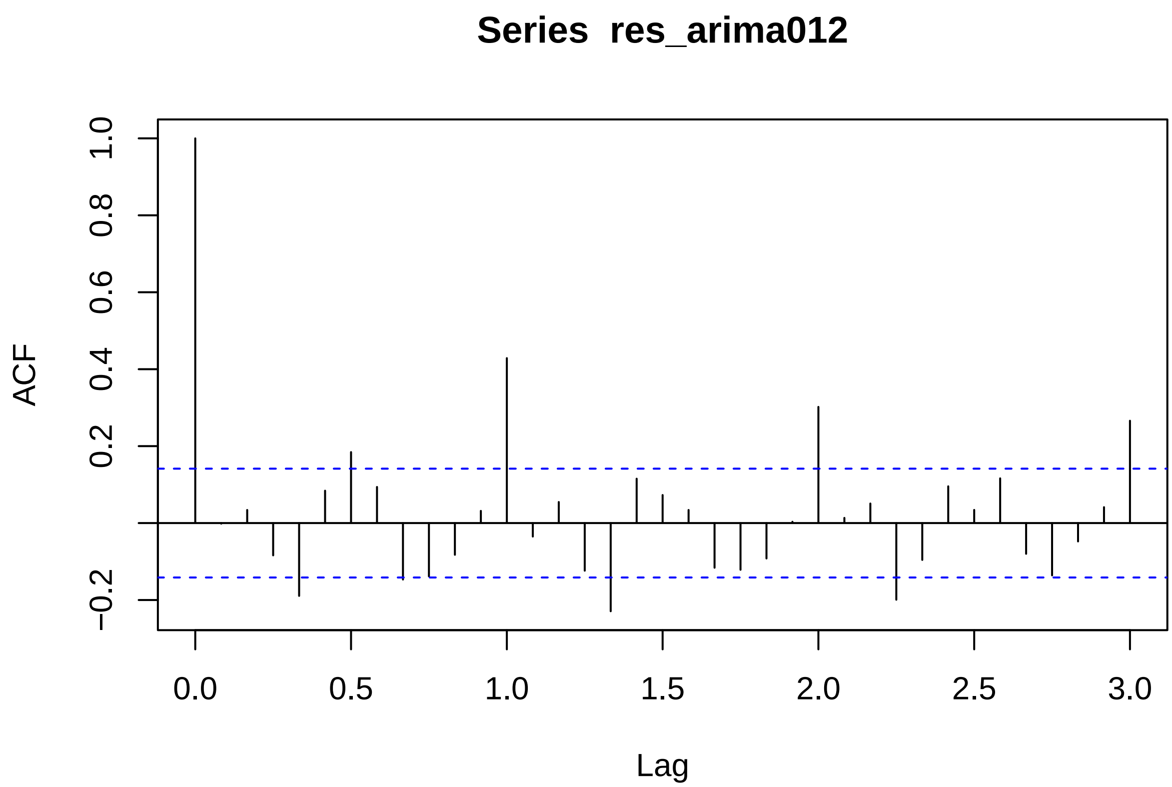
RMSE = 3441.309. Try an MA(2) or MA(4) model and address the seasonality later.

ARIMA(0,1,2) with constant

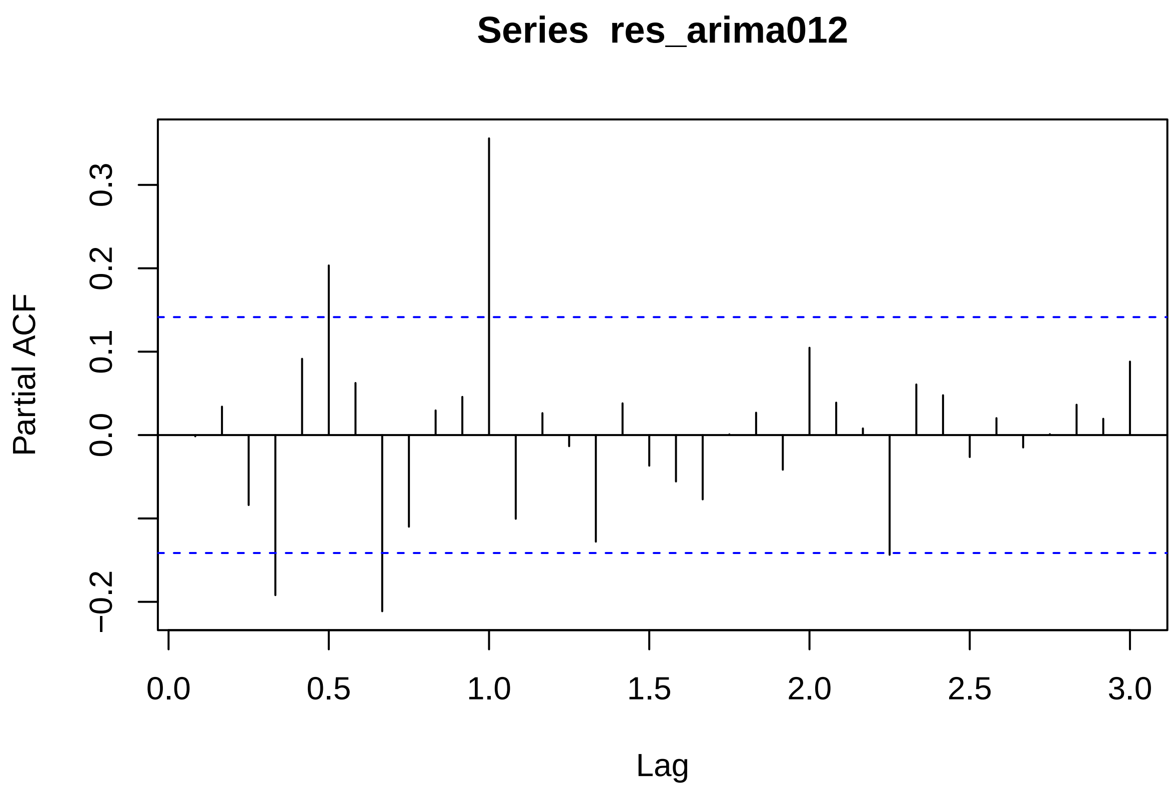
$ARIMA(0, 1, 2)$

```
man_arima012_fit_cons <- man_ts3 %>%
  model(arima012_constant = ARIMA(total_waste ~ 1 + pdq(0,1,2)))

zoo_arima012_fit <- arima(DSNY_MN_zoo_ts, order = 1 + c(0,1,2))
res_arima012 <- zoo_arima012_fit$residuals
acf(res_arima012, lag.max = 36)
```




```r
pacf(res_arima012, lag.max = 36)
```




```
accuracy(man_arima012_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 2586.
```

RMSE = 2586.395. We continue to see the RMSE decrease. Lag 4 in both the ACF and PACF plot is significant. We could try an ARIMA(0,1,4), but let us first address the seasonality on the current model.

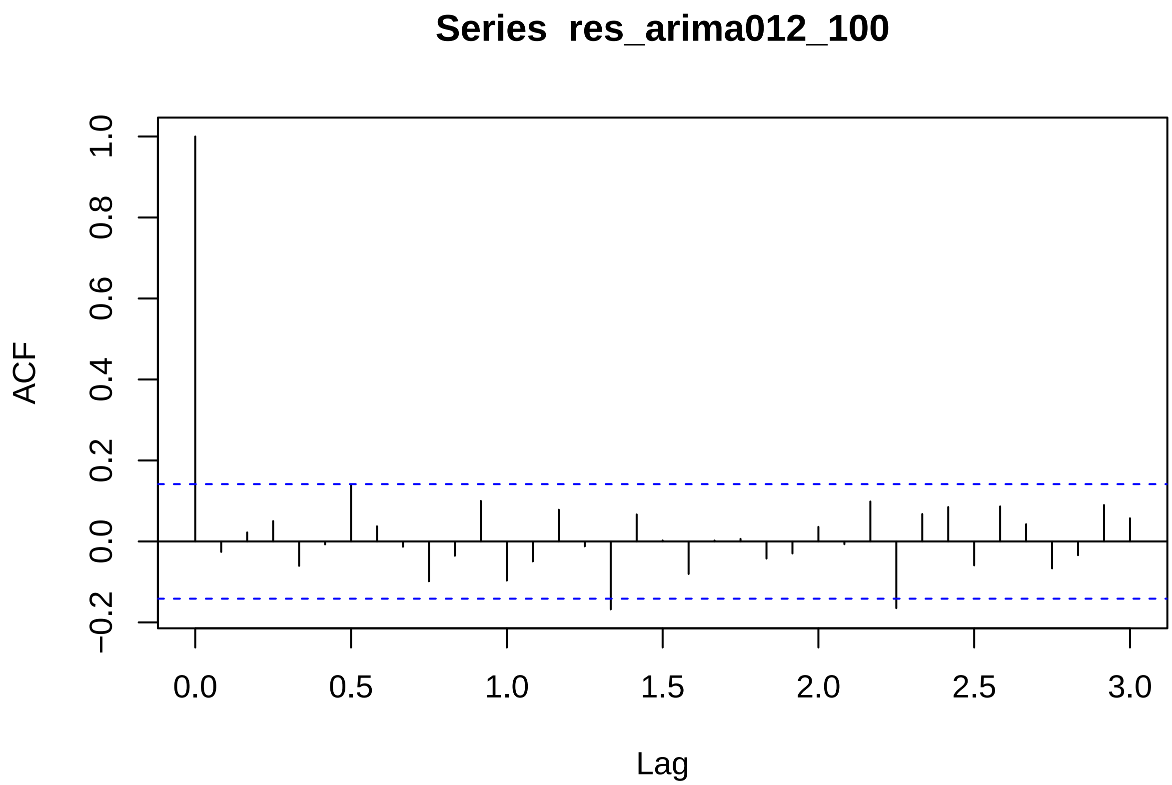
ARIMA(0,1,2)(1,0,0) with constant and seasonality

$ARIMA(0, 1, 2)(1, 0, 0)[12]$

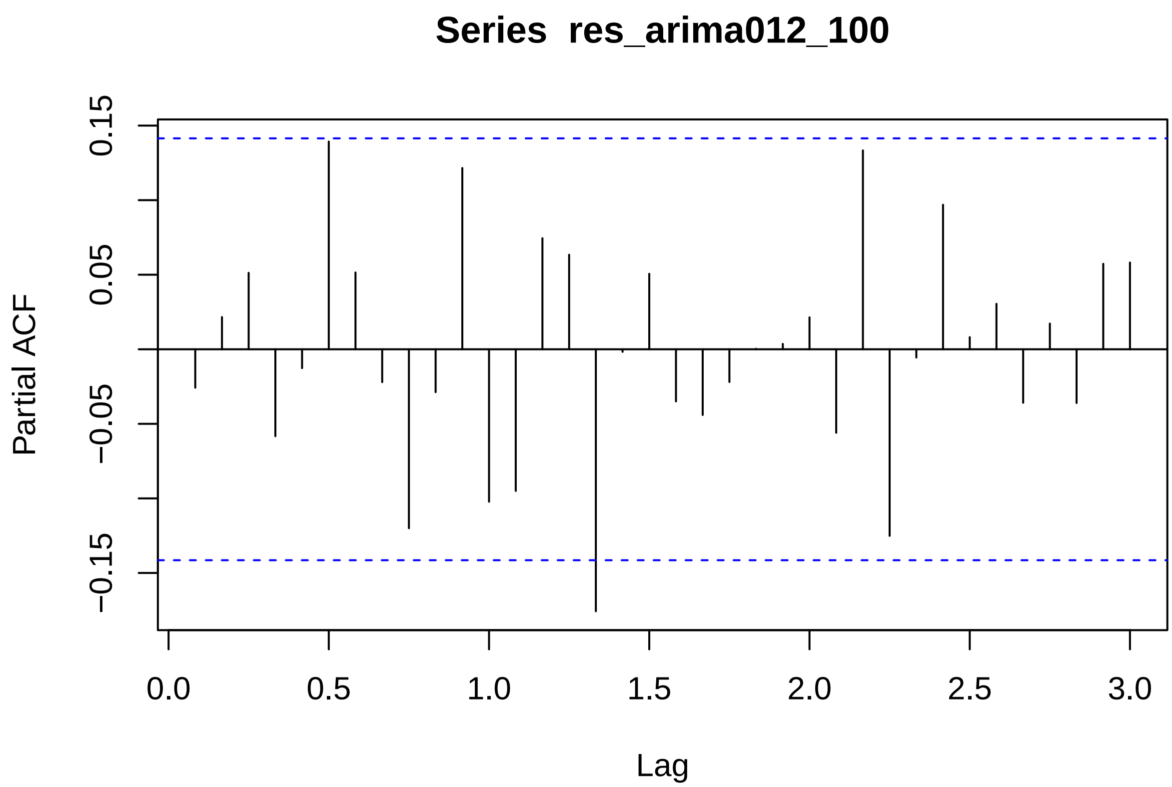
```
man_arima012_100_fit_cons <- man_ts3 %>%
  model(arima012_100_constant = ARIMA(total_waste ~ 1 +
                                       pdq(0,1,2) +
                                       PDQ(1,0,0,
                                           period = 12)))

zoo_arima012_100_fit <- arima(DSNY_MN_zoo_ts,
                              order = 1 + c(0,1,2),
                              seasonal = list(order = c(1,0L,0L),
                                              period = 12))
```

```r
res_arima012_100 <- zoo_arima012_100_fit$residuals
acf(res_arima012_100, lag.max = 36)
```




```r
pacf(res_arima012_100, lag.max = 36)
```




```
accuracy(man_arima012_100_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 2215.
```

RMSE = 2215.12. The majority of the ACF and PACF lag values are contained within bounds. However, lag 16 in both plots are significant. Being bounded b/w (-0.15, 0.15). There wouldn't be a direct way to address this lag without adding a high MA argument and potentialy overfitting this model.

Before turning to the auto arima model, lets work on a $ARIMA(0,1,4)(1,0,0)[12]$ model.

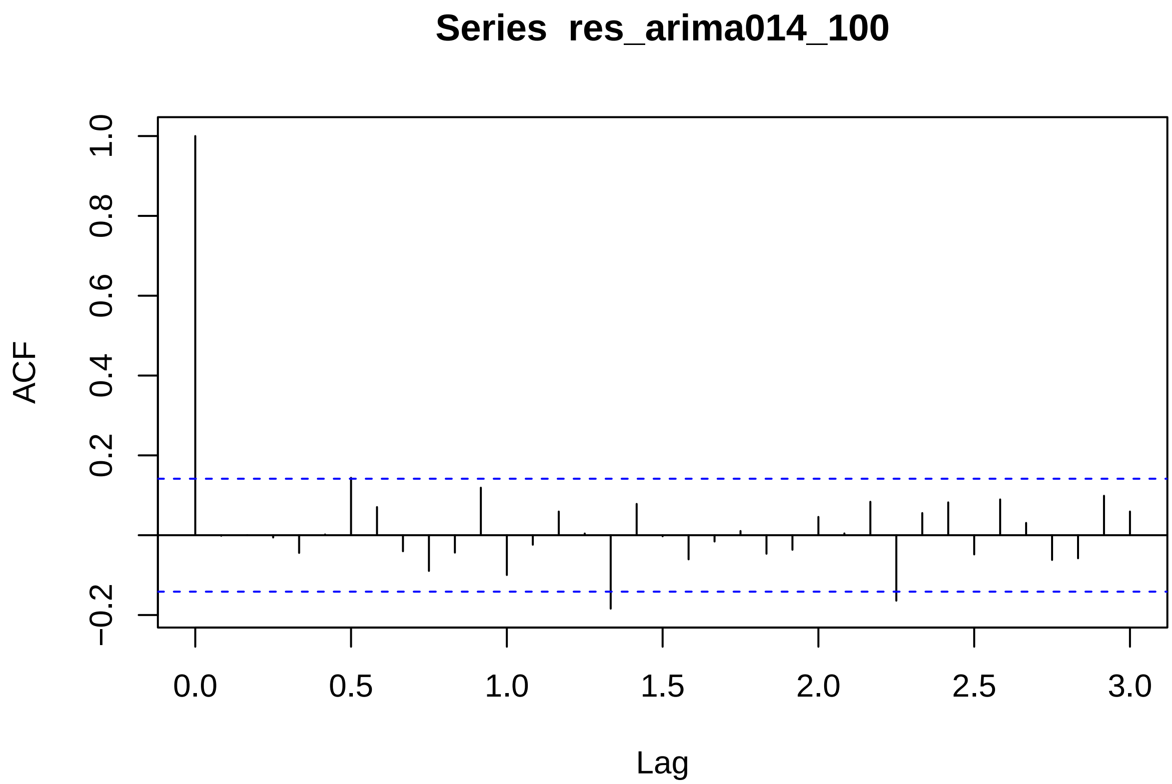
ARIMA(0,1,4)(1,0,0) with constant and seasonality

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

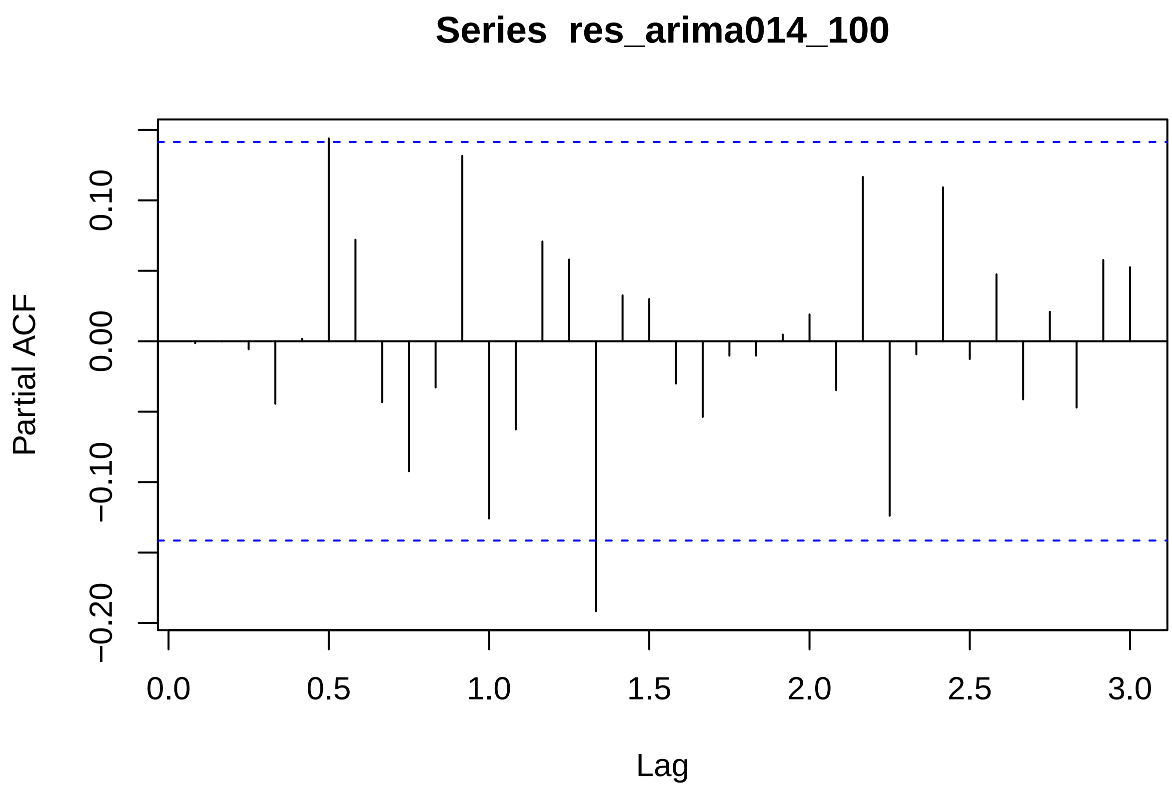
```
man_arima014_100_fit_cons <- man_ts3 %>%
  model(arima014_100_constant = ARIMA(total_waste ~ 1 +
                                      pdq(0,1,4) +
                                      PDQ(1,0,0,
                                          period = 12)))

zoo_arima014_100_fit <- arima(DSNY_MN_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(man_arima014_100_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 2208.
```

RMSE = 2208.185. Which is a decrease when compared to $ARIMA(0,1,2)(1,0,0)[12]$. Again, the majority of the lags are contained within the bounds.However we are not able to make the lag 12 insignificant in both plots.

Auto-arima

For our final model, we will look and compare the results of an auto-arima model from the feasts package.

```
man_auto_arima_fit_cons <- man_ts3 %>%
  model(stepwise = ARIMA(total_waste),
        search = ARIMA(total_waste,
                       stepwise = FALSE,
                       approximation = FALSE))

# zoo_arima014_100_fit <- arima(DSNY_MN_zoo_ts,
#                               order = 1 + c(0,1,4),
#                               seasonal = list(order = c(1,0L,0L),
```

```r
#                                                 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(man_auto_arima_fit_cons)[1:4]
```

```
## # A tibble: 2 x 4
##   .model   .type       ME  RMSE
##   <chr>    <chr>    <dbl> <dbl>
## 1 stepwise Training  17.4 2589.
## 2 search   Training  38.8 2450.
```

```r
#man_auto_arima_fit_cons %>% accuracy()
```

The stepwise model has RMSE = 2588.702, while the search model has RMSE = 2449.647. We will take a look at the ACF and PACF plots of the search model

Return the coefficients of the models above

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

```
## Series: total_waste
## Model: ARIMA(0,1,1) w/ drift
##
## Coefficients:
##           ma1  constant
##       -0.9140  -45.1428
## s.e.   0.0247   17.1493
##
## sigma^2 estimated as 6807750:  log likelihood=-1773.47
## AIC=3552.94   AICc=3553.07   BIC=3562.7
```

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

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

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

```
## Series: total_waste
## Model: ARIMA(0,1,5) w/ drift
##
## Coefficients:
##           ma1     ma2      ma3      ma4     ma5  constant
##       -0.9074  0.0146  -0.0520  -0.2763  0.3286  -50.0732
## s.e.   0.0688  0.0935   0.0834   0.0919  0.0648   19.6900
##
## sigma^2 estimated as 6227826:  log likelihood=-1763.23
## AIC=3540.47   AICc=3541.08   BIC=3563.23
```

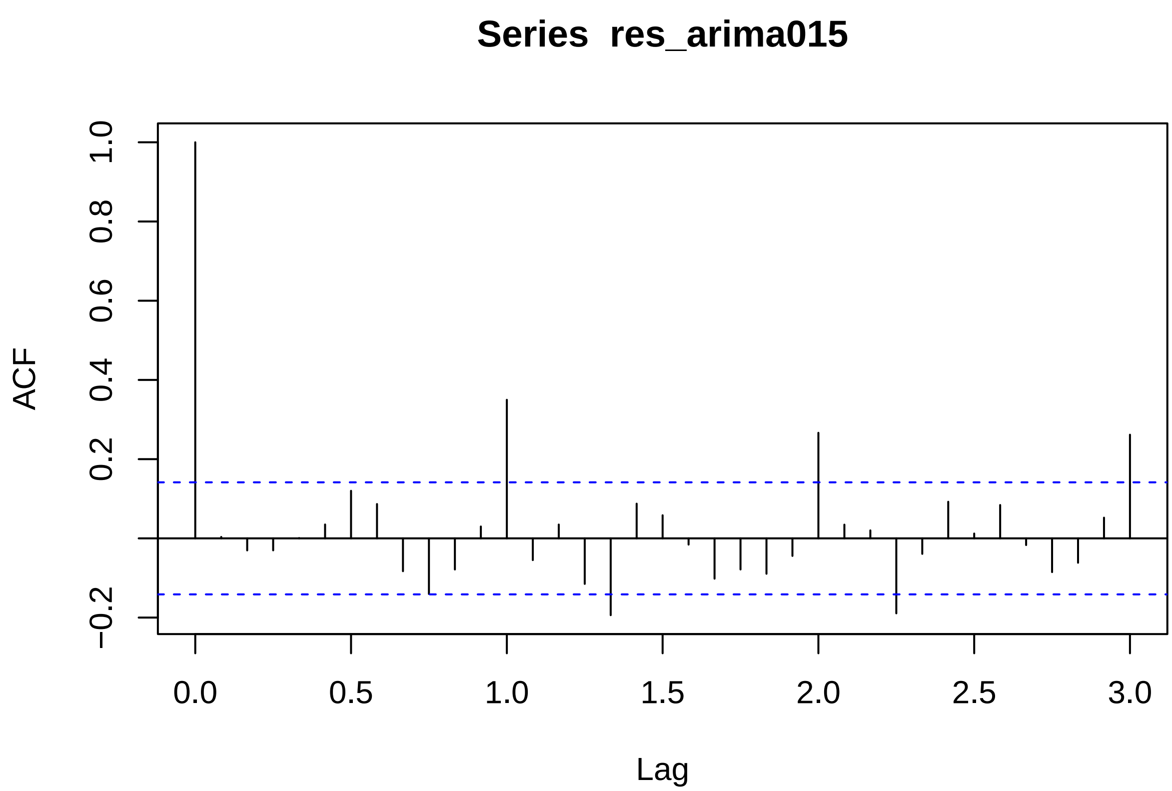
ARIMA(0,1,5) with constant from auto-arima

$ARIMA(0, 1, 5)$

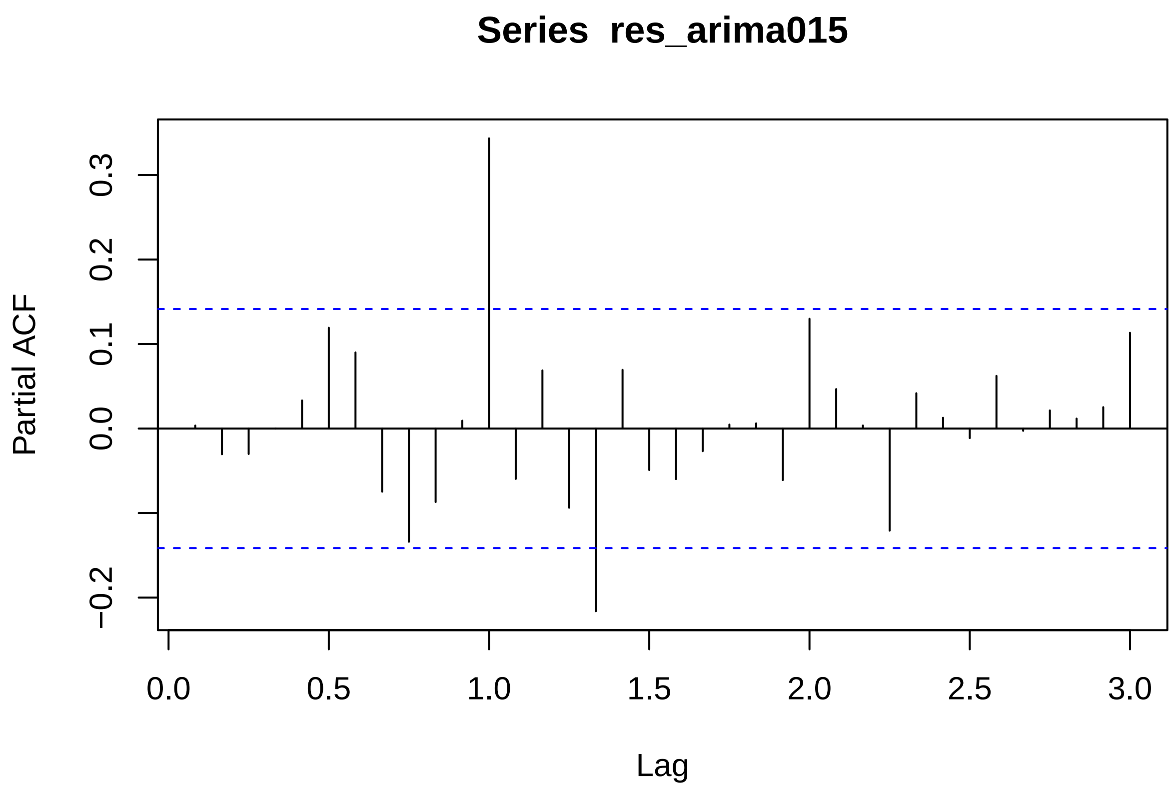
```
man_arima015_fit_cons <- man_ts3 %>%
  model(arima015_constant = ARIMA(total_waste ~ 1 +
                                      pdq(0,1,5)))

zoo_arima015_fit <- arima(DSNY_MN_zoo_ts,
                              order = 1 + c(0,1,5))

res_arima015 <- zoo_arima015_fit$residuals
acf(res_arima015, lag.max = 36)
```




```
pacf(res_arima015, lag.max = 36)
```




```r
accuracy(man_arima015_fit_cons)[4]
```

```
## # A tibble: 1 x 1
##    RMSE
##   <dbl>
## 1 2450.
```

Yeah this model was going to have significant lags without adressing the seasonality.

Summary of Models

$ARIMA(0,0,4)(1,0,0)_{12}$ has RMSE = 2386.46
$ARIMA(0,1,4)(1,0,0)[12]$ has RMSE = 2208.185
$ARIMA(0,1,2)(1,0,0)[12]$ has RMSE = 2215.12
$ARIMA(0,1,5)$ is the search model and has RMSE = 2449.647

Plots and Visualizations

Preliminary forecast of $ARIMA(0,0,4)(1,0,0)_{12}$