DATA624: Project 1

Donald Butler

2023-10-29

Contents

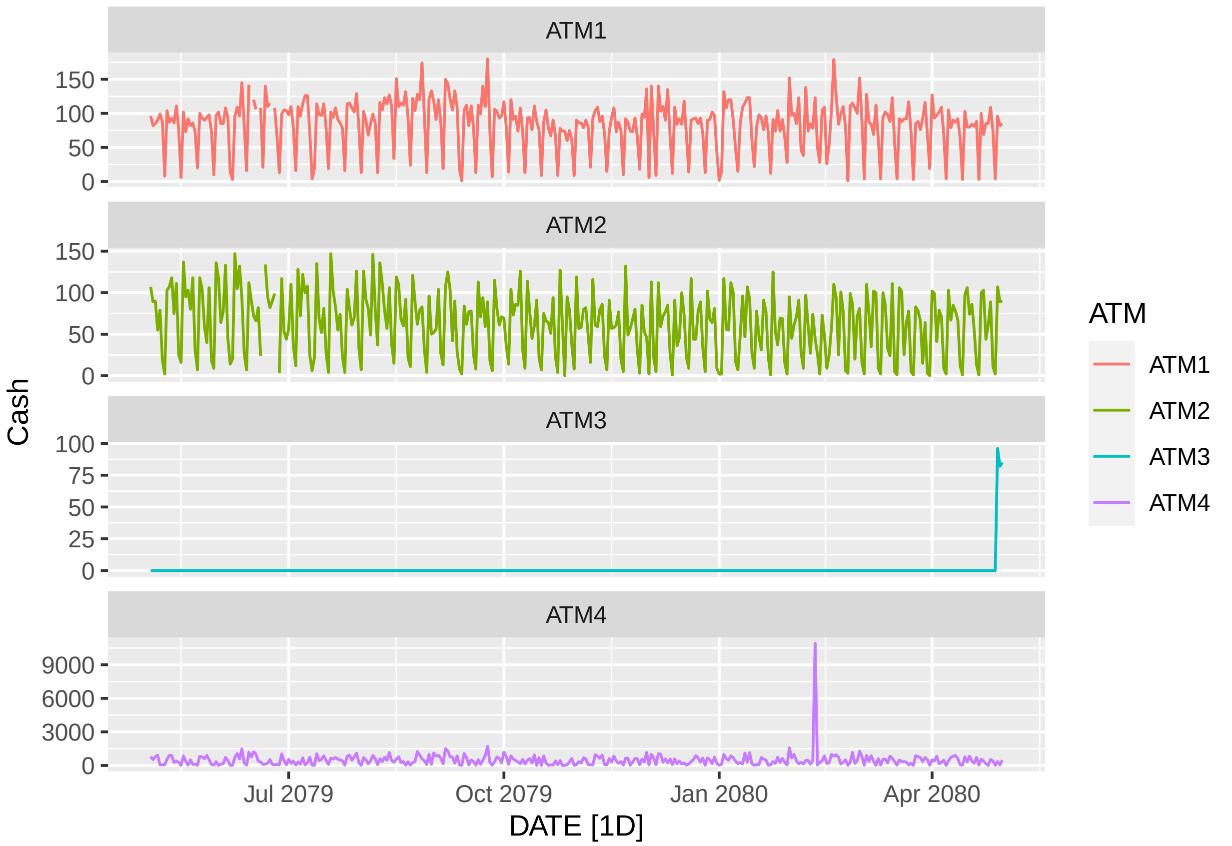
Part A

ATM Forecast - *ATM624Data.xlsx*

In part A, I want you to forecast how much cash is taken out of 4 different ATM machines for May 2010. The data is given in a single file. The variable 'Cash' is provided in hundreds of dollars, other than that it is straight forward. I am being somewhat ambiguous on purpose to make this have a little more business feeling. Explain and demonstrate your process, techniques used and not used, and your actual forecast. I am giving you data via an excel file, please provide your written report on your findings, visuals, discussion and your R code via an RPubs link along with the actual rmd file. Also please submit the forecast which you will put in an Excel readable file.

Loading Data

Loaded data from an Excel file into a `tsibble` object. The DATE column needed to be converted from an Excel datetime value, to a date type. Some rows in the file were missing values for the ATM value, so those were filtered out.



Looking at the data, we see two anomalies that will need to be addressed. ATM3 appears to be a newly installed machine, and only has a few days of data. ATM4 has a single value far outside the normal range of values.

Each series is following a weekly trend, with Thursdays having the lowest ATM usage.

Missing Values

A quick check for missing values shows that there are a total of 5 missing values.

```
ATM |>
  filter(is.na(Cash))
```

```
## # A tsibble: 5 x 3 [1D]
## # Key:       ATM [2]
##   DATE       ATM    Cash
##   <date>     <chr> <dbl>
## 1 2079-06-15 ATM1     NA
## 2 2079-06-18 ATM1     NA
## 3 2079-06-24 ATM1     NA
## 4 2079-06-20 ATM2     NA
## 5 2079-06-26 ATM2     NA
```

To preserve the seasonality of the data, we will interpolate the missing values using an ARIMA model.

```r
ATM <- ATM |>
  model(ARIMA(Cash)) |>
  interpolate(ATM)

ATM |>
  filter((ATM == 'ATM1' & (DATE == '2009-06-13' | DATE == '2009-06-16' | DATE == '2009-06-22')) | (ATM =
```

```
## # A tsibble: 0 x 3 [?]
## # Key:       ATM [0]
## # i 3 variables: ATM <chr>, DATE <date>, Cash <dbl>
```

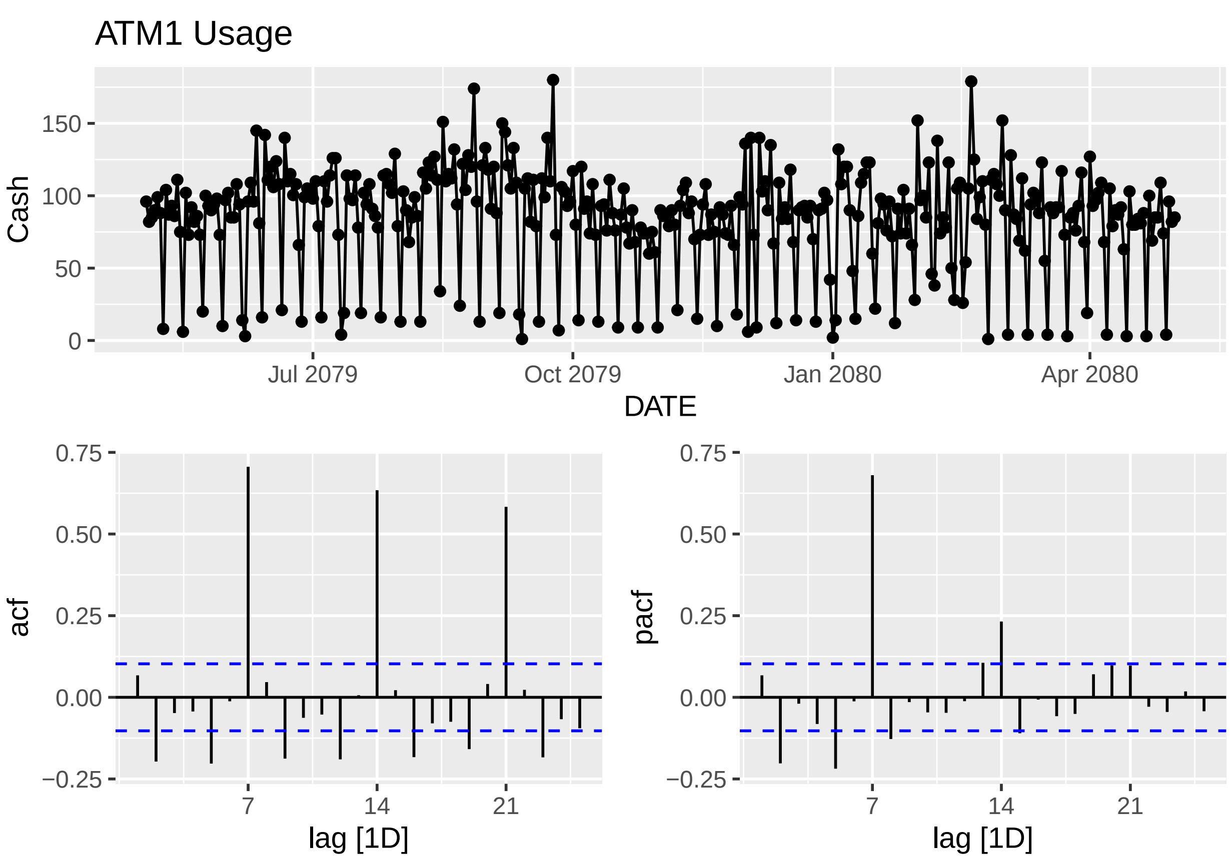
ATM Models

ATM1

```r
ATM1 <- ATM |>
  filter(ATM == 'ATM1')
```

Series Exploration The data clearly indicates a weekly seasonal component.

```r
ATM1 |>
  gg_tsdisplay(Cash, plot_type = 'partial') +
  labs(title = 'ATM1 Usage')
```



Transformation Identify a lambda value for a Box-Cox transformation.

```
lambda <- ATM1 |>
  features(Cash, features = guerrero) |>
  pull(lambda_guerrero)

lambda
```

```
## [1] 0.2622969
```

Models We will construct a few models to determine the best choice.

```
fit1 <- ATM1 |>
  model(
    additive = ETS(box_cox(Cash, lambda) ~ error('A') + trend('N') + season('A')),
    multiplicative = ETS(box_cox(Cash, lambda) ~ error('M') + trend('N') + season('M')),
    arima = ARIMA(box_cox(Cash, lambda), stepwise = FALSE)
  )

fit1 |>
  glance() |>
  select(.model:BIC) |>
  arrange(AIC)
```

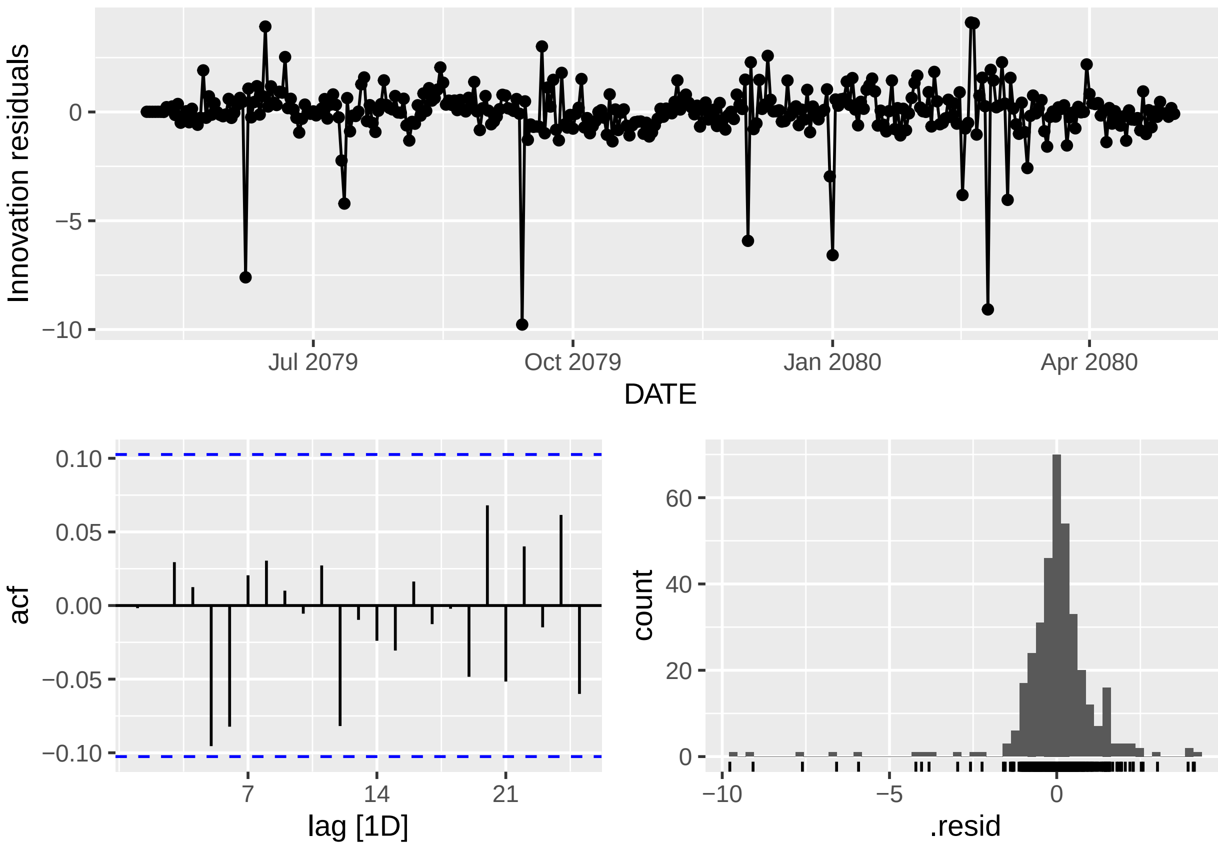
```
## # A tibble: 3 x 6
##   .model         sigma2 log_lik   AIC  AICc   BIC
##   <chr>           <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima          1.77     -611. 1229. 1229. 1245.
## 2 additive       1.81    -1180. 2381. 2381. 2420.
## 3 multiplicative 0.0383  -1220. 2459. 2460. 2498.
```

The ARIMA model has the lowest AIC statistic.

```
fit1 |>
  select(arima) |>
  report()
```

```
## Series: Cash
## Model: ARIMA(0,0,2)(0,1,1)[7]
## Transformation: box_cox(Cash, lambda)
##
## Coefficients:
##          ma1      ma2     sma1
##       0.1105  -0.1088  -0.6419
## s.e.  0.0524   0.0521   0.0431
##
## sigma^2 estimated as 1.771:  log likelihood=-610.69
## AIC=1229.38   AICc=1229.49   BIC=1244.9
```

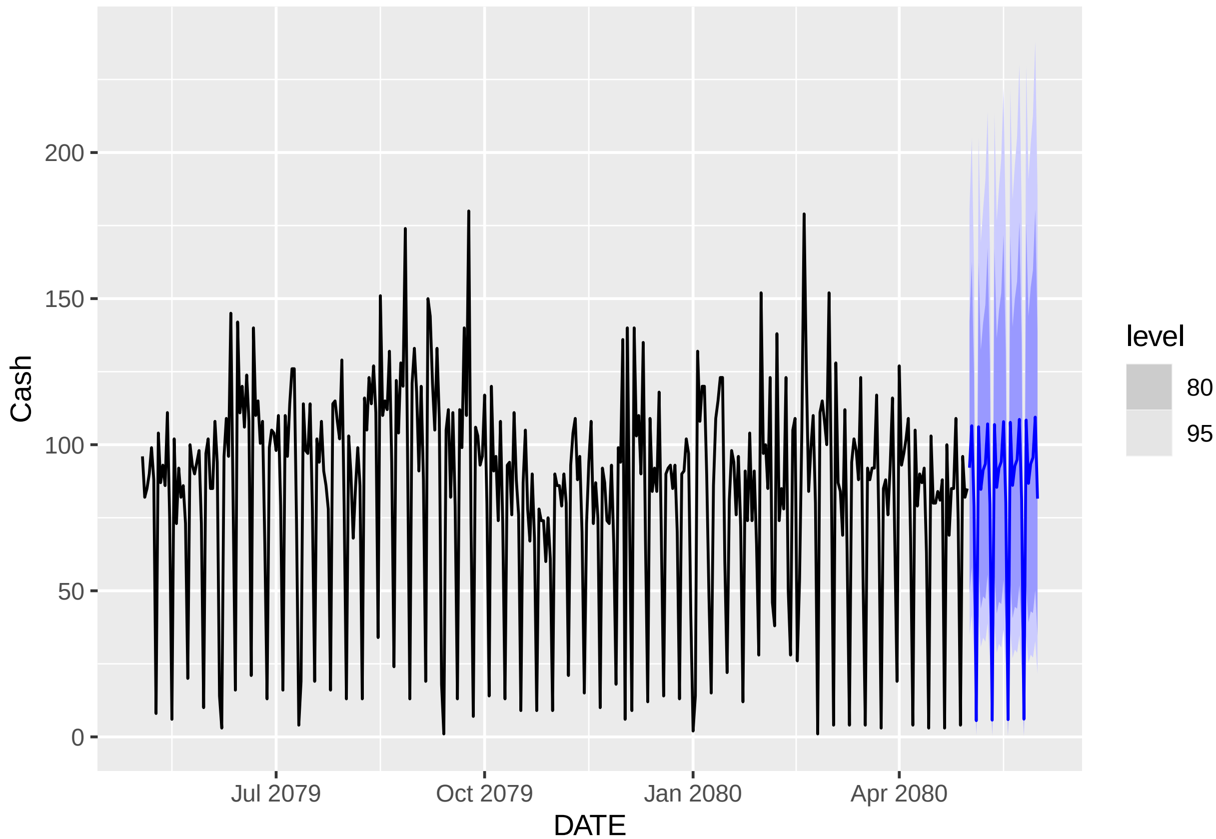
```
fit1 |>
  select(arima) |>
  gg_tsresiduals()
```



The residuals for the ARIMA model appear to normal centered around zero and have no spikes outside of the confidence interval.

```
fc1 <- fit1 |>
  select(ATM, arima) |>
  forecast(h = 31)
```

```
fc1 |>
  autoplot(ATM1)
```



Forecast

```
fc1 |>
  as.data.frame() |>
  select(DATE, .mean) |>
  rename(Cash = .mean) |>
  mutate(Cash = round(Cash,0))
```

```
##          DATE Cash
## 1  2080-05-02   92
## 2  2080-05-03  106
## 3  2080-05-04   79
## 4  2080-05-05    6
## 5  2080-05-06  106
## 6  2080-05-07   85
## 7  2080-05-08   91
## 8  2080-05-09   93
## 9  2080-05-10  107
## 10 2080-05-11   80
## 11 2080-05-12    6
## 12 2080-05-13  107
## 13 2080-05-14   85
## 14 2080-05-15   92
## 15 2080-05-16   94
## 16 2080-05-17  108
## 17 2080-05-18   80
## 18 2080-05-19    6
## 19 2080-05-20  108
## 20 2080-05-21   86
```

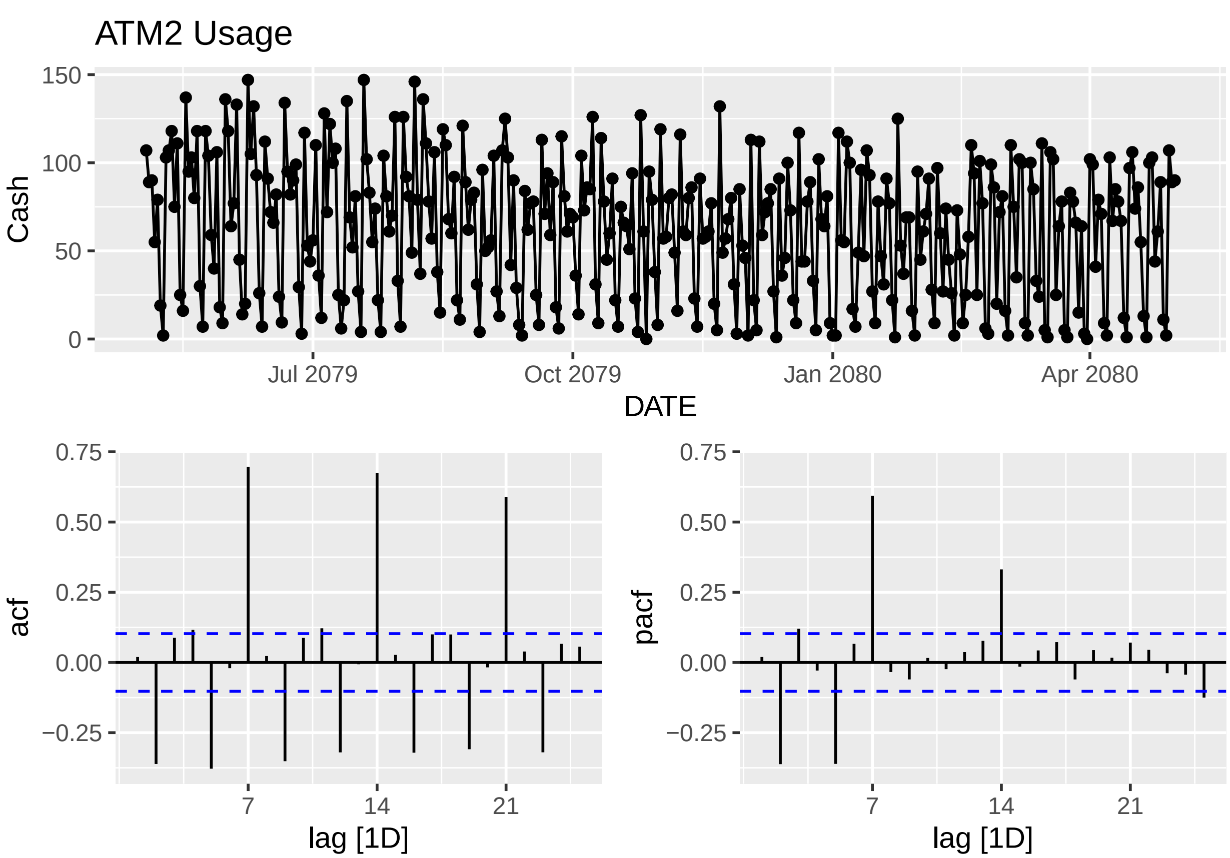
```
## 21 2080-05-22   93
## 22 2080-05-23   95
## 23 2080-05-24  109
## 24 2080-05-25   81
## 25 2080-05-26    6
## 26 2080-05-27  108
## 27 2080-05-28   87
## 28 2080-05-29   93
## 29 2080-05-30   96
## 30 2080-05-31  109
## 31 2080-06-01   82
```

ATM2

```r
ATM2 <- ATM |>
  filter(ATM == 'ATM2')
```

Series Exploration The data clearly indicates a weekly seasonal component with the spikes in the ACF at 7, 14, and 21.

```r
ATM2 |>
  gg_tsdisplay(Cash, plot_type = 'partial') +
  labs(title = 'ATM2 Usage')
```



Transformation Identify a lambda value for a Box-Cox transformation.

```
lambda <- ATM2 |>
  features(Cash, features = guerrero) |>
  pull(lambda_guerrero)

lambda
```

```
## [1] 0.6746523
```

Models We will construct a few models to determine the best choice.

```
fit2 <- ATM2 |>
  model(
    additive = ETS(box_cox(Cash, lambda) ~ error('A') + trend('N') + season('A')),
    multiplicative = ETS(box_cox(Cash, lambda) ~ error('M') + trend('N') + season('M')),
    arima = ARIMA(box_cox(Cash, lambda), stepwise = FALSE)
  )

fit2 |>
  glance() |>
  select(.model:BIC) |>
  arrange(AIC)
```

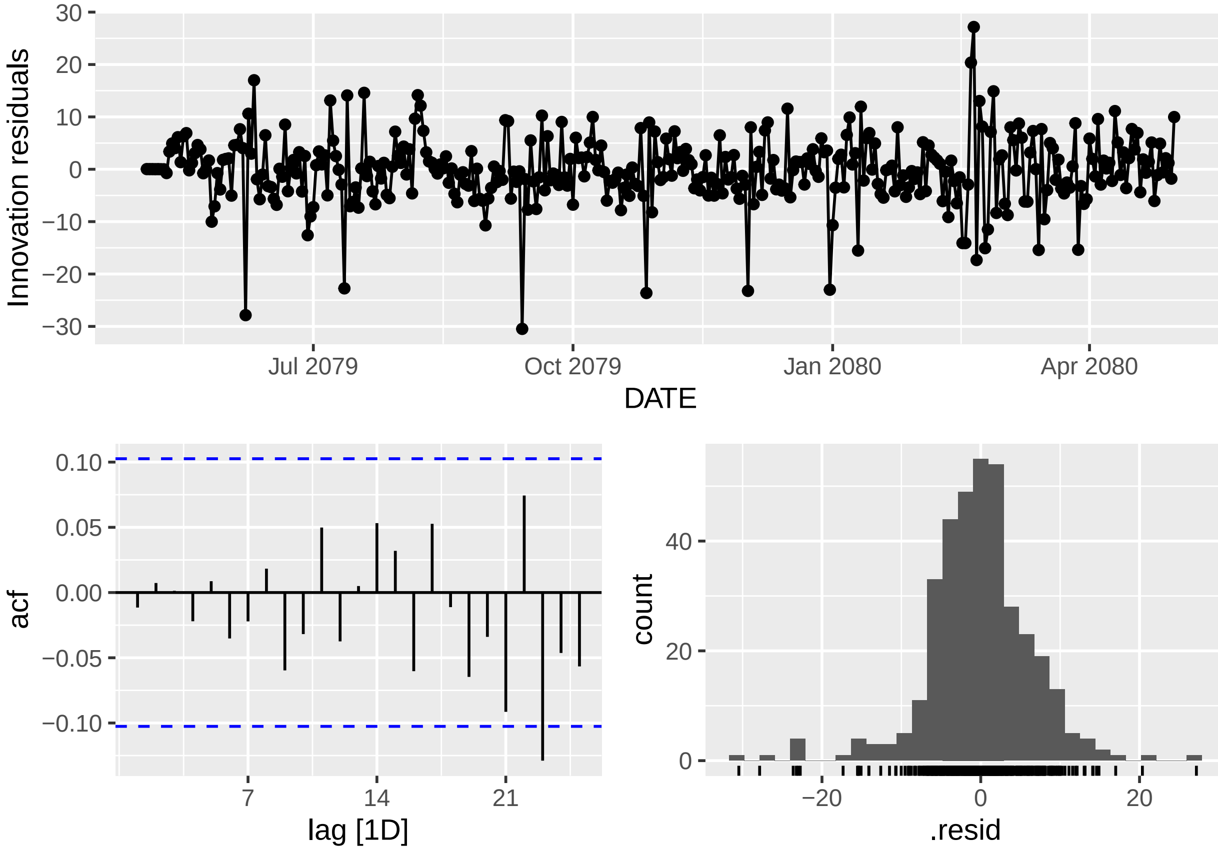
```
## # A tibble: 3 x 6
##   .model         sigma2 log_lik   AIC  AICc   BIC
##   <chr>           <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima          44.9    -1188. 2390. 2391. 2418.
## 2 additive       47.5    -1777. 3573. 3574. 3612.
## 3 multiplicative  0.245  -1932. 3885. 3886. 3924.
```

The ARIMA model has the lowest AIC statistic.

```
fit2 |>
  select(arima) |>
  report()
```

```
## Series: Cash
## Model: ARIMA(5,0,0)(0,1,1)[7]
## Transformation: box_cox(Cash, lambda)
##
## Coefficients:
##          ar1      ar2     ar3     ar4      ar5     sma1
##       0.0540  -0.1185  0.0037  0.0909  -0.2083  -0.6671
## s.e.  0.0516   0.0545  0.0519  0.0514   0.0538   0.0438
##
## sigma^2 estimated as 44.94:  log likelihood=-1188.17
## AIC=2390.35   AICc=2390.67   BIC=2417.51
```

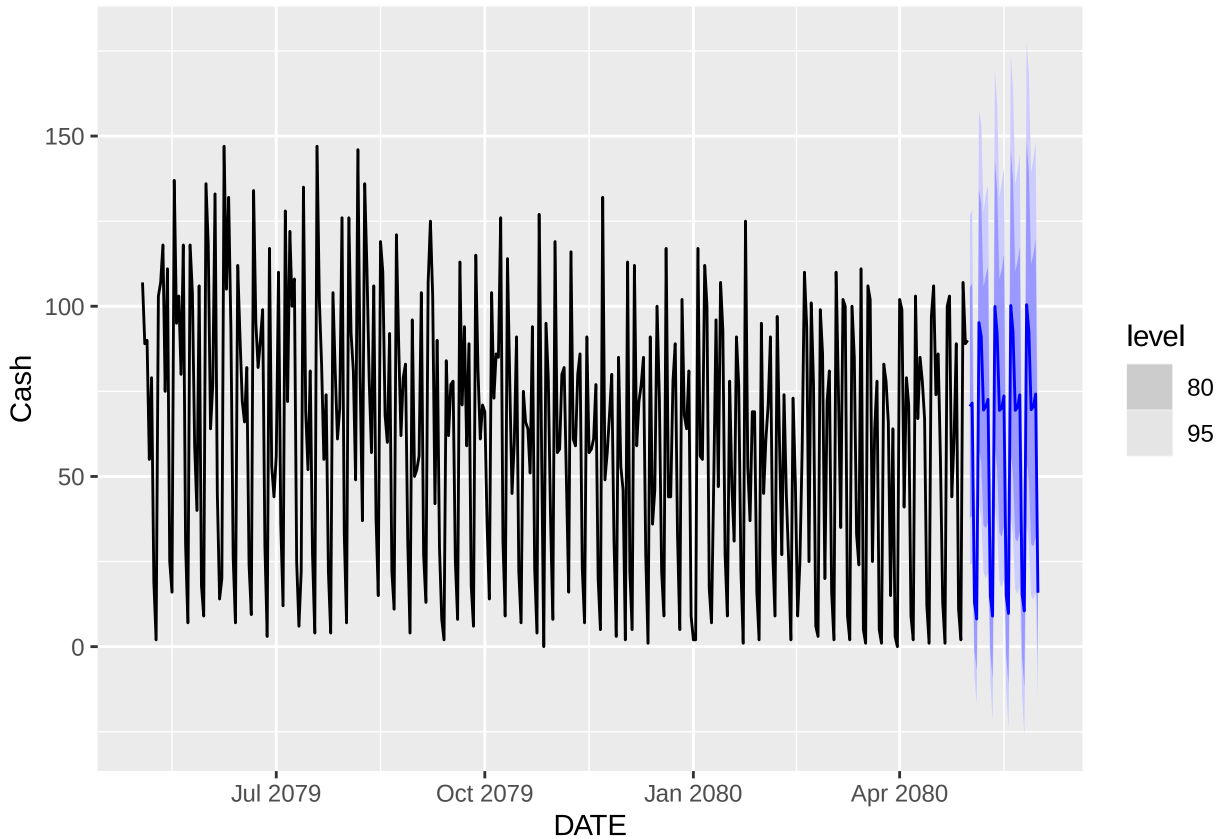
```
fit2 |>
  select(arima) |>
  gg_tsresiduals()
```



The residuals for the ARIMA model appear to normal centered around zero and have no spikes outside of the confidence interval.

```
fc2 <- fit2 |>
  select(ATM, arima) |>
  forecast(h = 31)
```

```
fc2 |>
  autoplot(ATM2)
```



Forecast

```
fc2 |>
  as.data.frame() |>
  select(DATE, .mean) |>
  rename(Cash = .mean) |>
  mutate(Cash = round(Cash,0))
```

```
##           DATE Cash
## 1  2080-05-02   71
## 2  2080-05-03   72
## 3  2080-05-04   13
## 4  2080-05-05    8
## 5  2080-05-06   95
## 6  2080-05-07   91
## 7  2080-05-08   69
## 8  2080-05-09   71
## 9  2080-05-10   73
## 10 2080-05-11   15
## 11 2080-05-12    9
## 12 2080-05-13  100
## 13 2080-05-14   92
## 14 2080-05-15   69
## 15 2080-05-16   70
## 16 2080-05-17   74
## 17 2080-05-18   15
## 18 2080-05-19   10
## 19 2080-05-20  100
## 20 2080-05-21   92
```

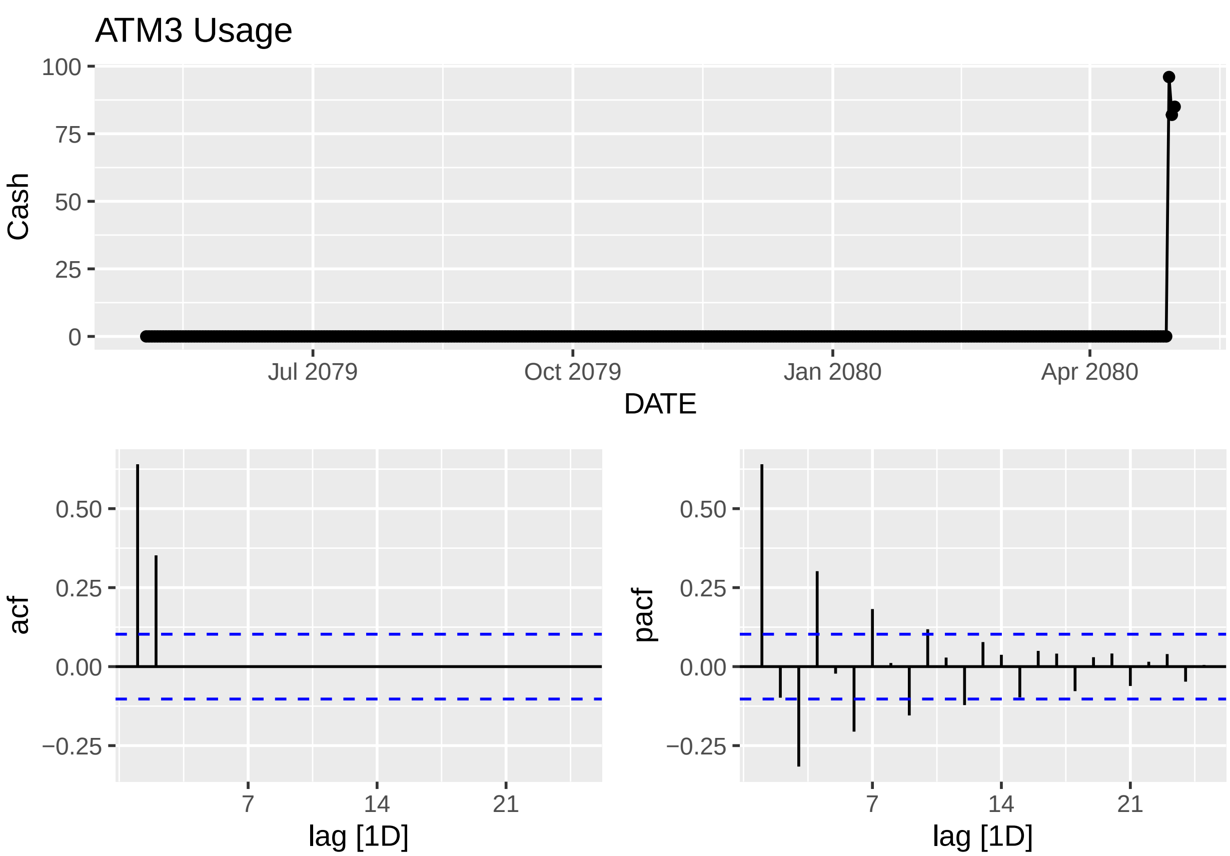
```
## 21 2080-05-22   69
## 22 2080-05-23   70
## 23 2080-05-24   74
## 24 2080-05-25   15
## 25 2080-05-26   10
## 26 2080-05-27  100
## 27 2080-05-28   93
## 28 2080-05-29   70
## 29 2080-05-30   71
## 30 2080-05-31   74
## 31 2080-06-01   16
```

ATM3

```
ATM3 <- ATM |>
  filter(ATM == 'ATM3')
```

Series Exploration The data for this ATM indicates that it is newly installed. While we may expect a similar weekly pattern to hold with this ATM, there isn't enough data to establish a pattern.

```
ATM3 |>
  gg_tsdisplay(Cash, plot_type = 'partial') +
  labs(title = 'ATM3 Usage')
```



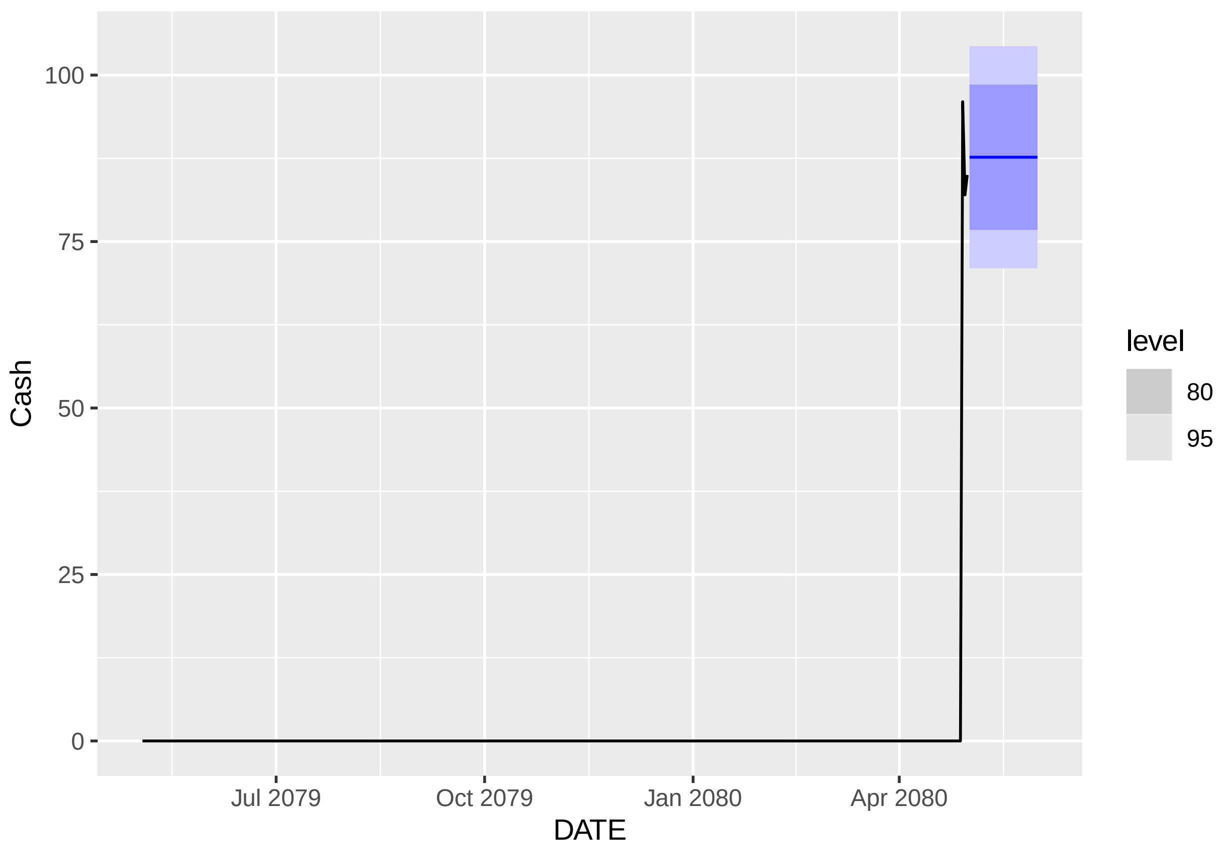
Models We will construct a model based on the MEAN of the 3 values in the data set.

```
fit3 <- ATM3 |>
  filter(Cash > 0) |>
  model(MEAN(Cash))

fit3 |>
  report()
```

```
## Series: Cash
## Model: MEAN
##
## Mean: 87.6667
## sigma^2: 54.3333
```

```
fc3 <- fit3 |>
  forecast(h = 31)
```

```
fc3 |>
  autoplot(ATM3)
```



Forecast

```
fc3 |>
  as.data.frame() |>
  select(DATE, .mean) |>
  rename(Cash = .mean) |>
  mutate(Cash = round(Cash,0))
```

```
##          DATE Cash
## 1  2080-05-02   88
## 2  2080-05-03   88
## 3  2080-05-04   88
## 4  2080-05-05   88
## 5  2080-05-06   88
## 6  2080-05-07   88
## 7  2080-05-08   88
## 8  2080-05-09   88
## 9  2080-05-10   88
## 10 2080-05-11   88
## 11 2080-05-12   88
## 12 2080-05-13   88
## 13 2080-05-14   88
## 14 2080-05-15   88
## 15 2080-05-16   88
## 16 2080-05-17   88
## 17 2080-05-18   88
## 18 2080-05-19   88
## 19 2080-05-20   88
## 20 2080-05-21   88
## 21 2080-05-22   88
## 22 2080-05-23   88
## 23 2080-05-24   88
## 24 2080-05-25   88
## 25 2080-05-26   88
## 26 2080-05-27   88
## 27 2080-05-28   88
## 28 2080-05-29   88
## 29 2080-05-30   88
## 30 2080-05-31   88
## 31 2080-06-01   88
```

ATM4

Outliers There is a clear outlier in ATM4 and it would be best to remove it before generating models on the data.

```
ATM |>
  filter(ATM == 'ATM4') |>
  mutate(mean = mean(Cash)) |>
  filter(Cash > 10 * mean)
```

```
## # A tsibble: 1 x 4 [1D]
## # Key:       ATM [1]
##   ATM   DATE         Cash  mean
```

```
##   <chr> <date>      <dbl> <dbl>
## 1 ATM4  2080-02-11 10920.  474.
```

We will replace this value with `NA`, then interpolate with an ARIMA model like was done with the missing values.

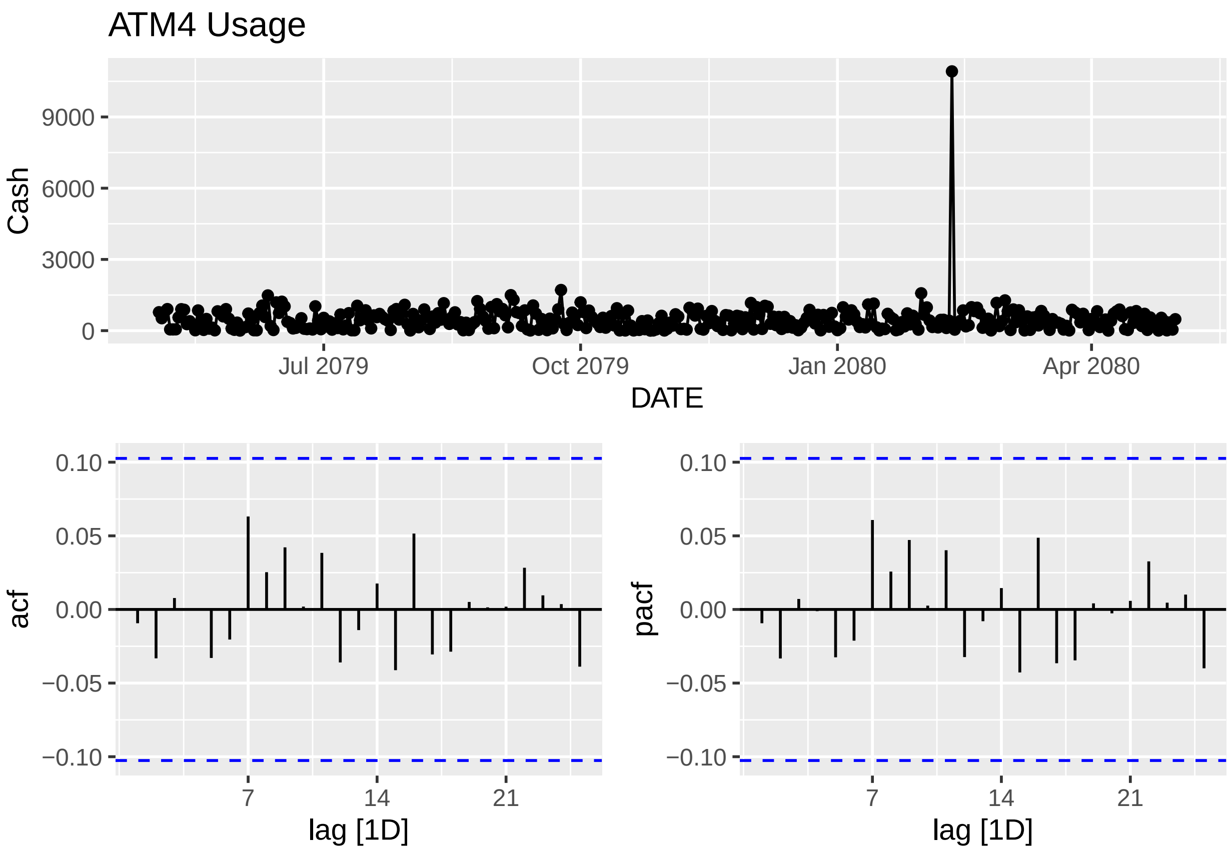
```
ATM <- ATM |>
  mutate(Cash = replace(Cash, ATM == 'ATM4' & DATE == '2010-02-09', NA))

ATM <- ATM |>
  model(ARIMA(Cash)) |>
  interpolate(ATM)

ATM |>
  filter(ATM == 'ATM4' & DATE == '2010-02-09')
```

```
## # A tsibble: 0 x 3 [?]
## # Key:       ATM [0]
## # i 3 variables: ATM <chr>, DATE <date>, Cash <dbl>
```

```
ATM4 <- ATM |>
  filter(ATM == 'ATM4')
```

Series Exploration The data appears to follow the same weekly pattern as ATM1 and ATM2.

```
ATM4 |>
  gg_tsdisplay(Cash, plot_type = 'partial') +
  labs(title = 'ATM4 Usage')
```



Transformation Identify a lambda value for a Box-Cox transformation.

```
lambda <- ATM4 |>
  features(Cash, features = guerrero) |>
  pull(lambda_guerrero)

lambda
```

```
## [1] -0.0737252
```

Models We will construct a few models to determine the best choice.

```
fit4 <- ATM4 |>
  model(
    additive = ETS(box_cox(Cash, lambda) ~ error('A') + trend('N') + season('A')),
    multiplicative = ETS(box_cox(Cash, lambda) ~ error('M') + trend('N') + season('M')),
    arima = ARIMA(box_cox(Cash, lambda), stepwise = FALSE)
  )

fit4 |>
  glance() |>
  select(.model:BIC) |>
  arrange(AIC)
```

```
## # A tibble: 3 x 6
```

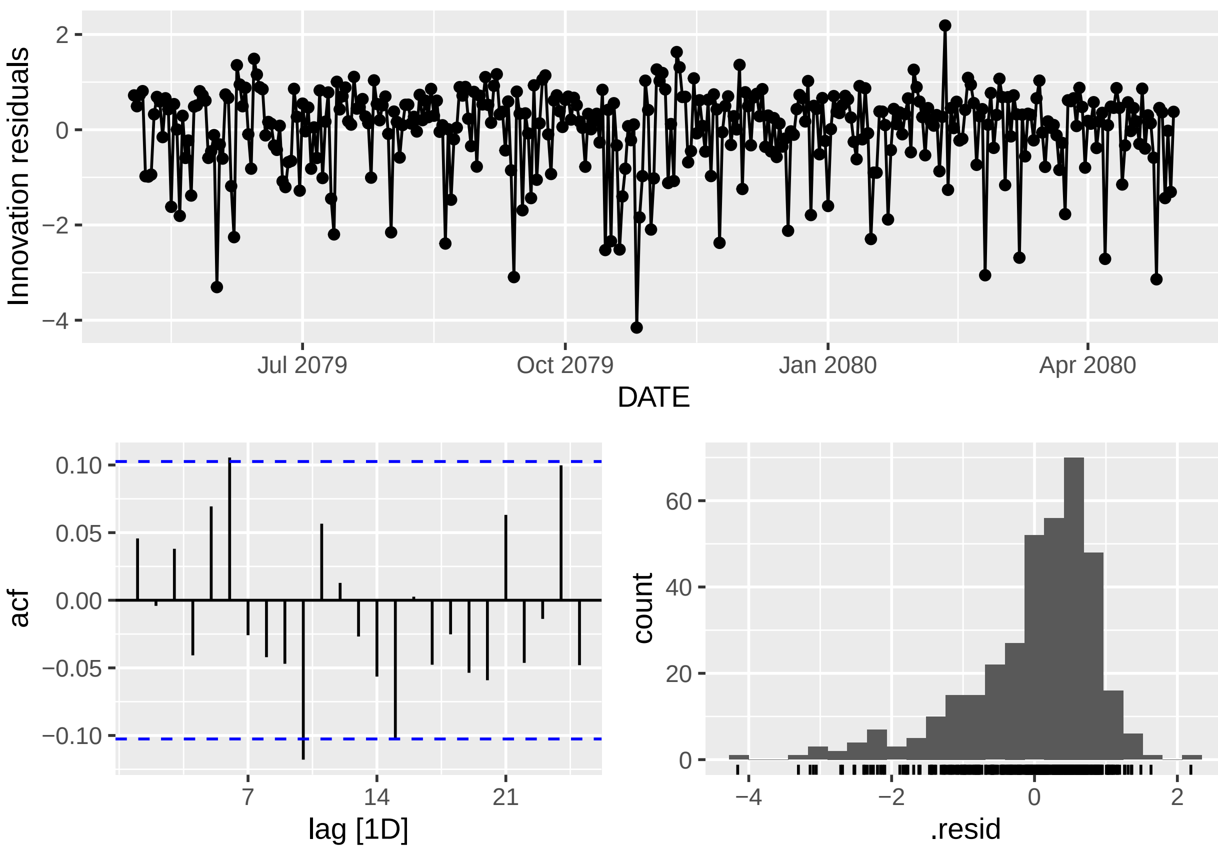
```
##   .model         sigma2 log_lik   AIC  AICc   BIC
##   <chr>           <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima          0.842    -486.  979.  979.  995.
## 2 additive       0.807   -1033. 2086. 2087. 2125.
## 3 multiplicative 0.0409  -1034. 2089. 2089. 2128.
```

The ARIMA model has the lowest AIC statistic.

```
fit4 |>
  select(arima) |>
  report()
```

```
## Series: Cash
## Model: ARIMA(0,0,0)(2,0,0)[7] w/ mean
## Transformation: box_cox(Cash, lambda)
##
## Coefficients:
##         sar1    sar2  constant
##       0.2487  0.1947    2.4972
## s.e.  0.0521  0.0525    0.0468
##
## sigma^2 estimated as 0.8418:  log likelihood=-485.59
## AIC=979.18   AICc=979.29   BIC=994.78
```

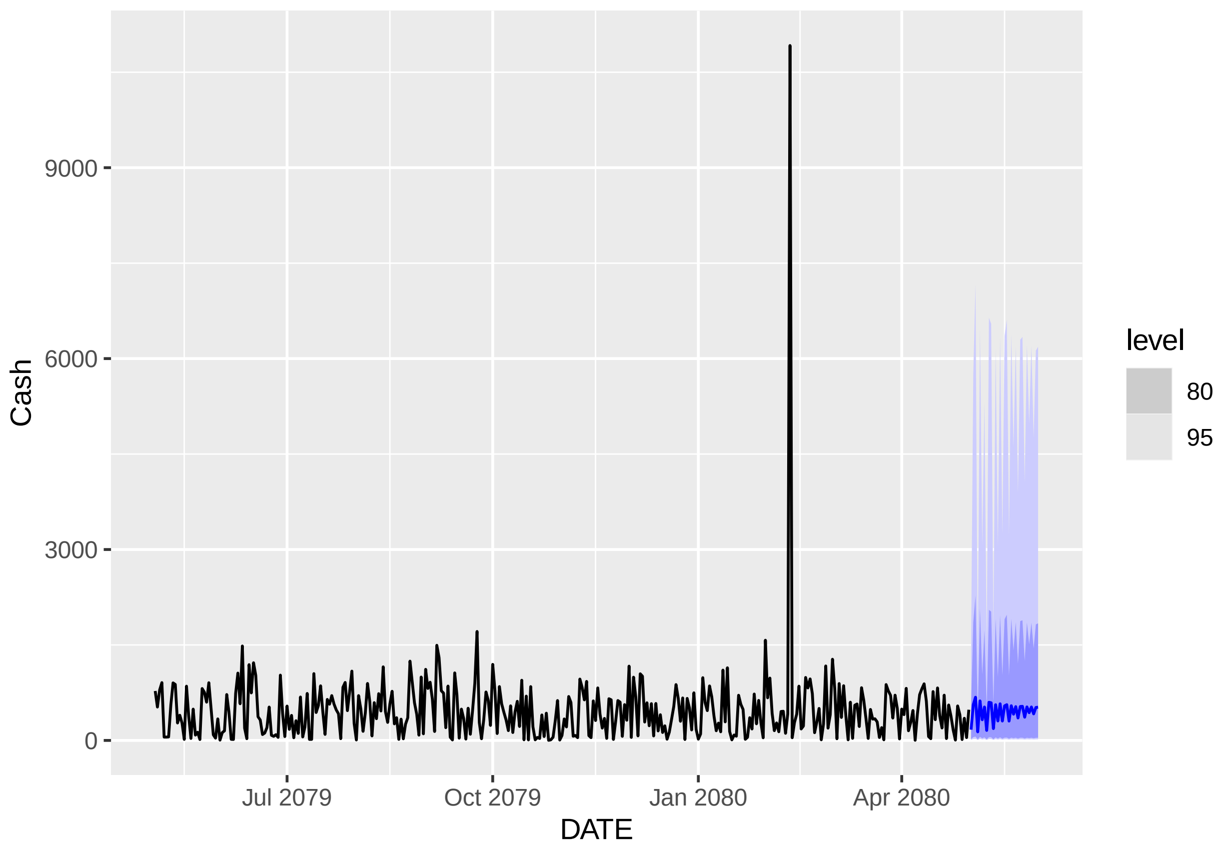
```
fit4 |>
  select(arima) |>
  gg_tsresiduals()
```



The residuals for the ARIMA model appear to normal centered around zero and have no spikes outside of the confidence interval.

```
fc4 <- fit4 |>
  select(ATM, arima) |>
  forecast(h = 31)
```

```
fc4 |>
  autoplot(ATM4)
```



Forecast

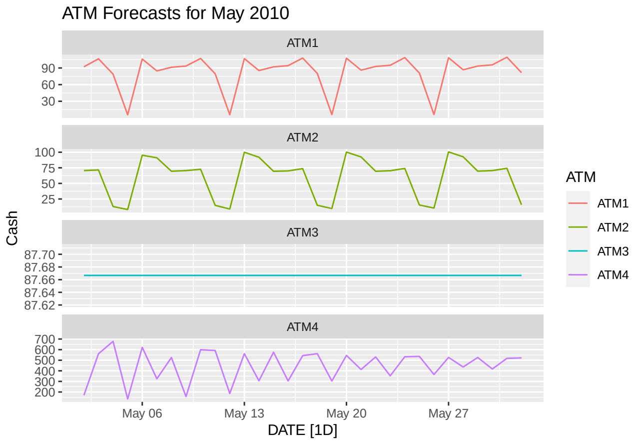
```
fc4 |>
  as.data.frame() |>
  select(DATE, .mean) |>
  rename(Cash = .mean) |>
  mutate(Cash = round(Cash,0))
```

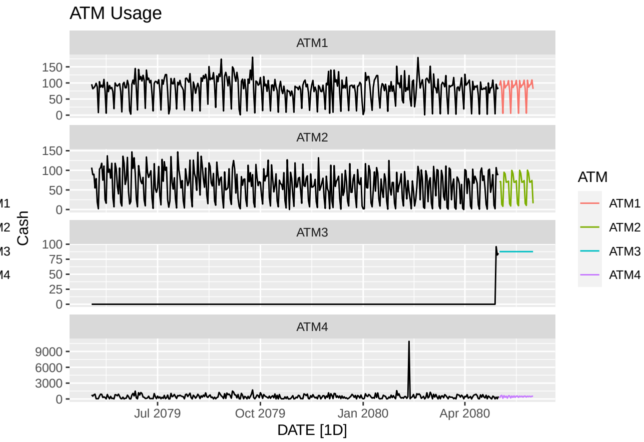
```
##          DATE Cash
## 1  2080-05-02  169
## 2  2080-05-03  561
## 3  2080-05-04  678
## 4  2080-05-05  134
```

```
## 5  2080-05-06  621
## 6  2080-05-07  325
## 7  2080-05-08  525
## 8  2080-05-09  157
## 9  2080-05-10  599
## 10 2080-05-11  592
## 11 2080-05-12  186
## 12 2080-05-13  562
## 13 2080-05-14  306
## 14 2080-05-15  576
## 15 2080-05-16  305
## 16 2080-05-17  543
## 17 2080-05-18  562
## 18 2080-05-19  304
## 19 2080-05-20  545
## 20 2080-05-21  412
## 21 2080-05-22  531
## 22 2080-05-23  353
## 23 2080-05-24  533
## 24 2080-05-25  536
## 25 2080-05-26  365
## 26 2080-05-27  527
## 27 2080-05-28  436
## 28 2080-05-29  526
## 29 2080-05-30  417
## 30 2080-05-31  518
## 31 2080-06-01  522
```

Forecasted Data

```r
fc <- fc1 |>
  bind_rows(fc2) |>
  bind_rows(fc3) |>
  bind_rows(fc4) |>
  as.data.frame() |>
  select(DATE, ATM, .mean) |>
  rename(Cash = .mean)

fc |> as_tsibble(index = DATE, key = ATM) |>
  autoplot(Cash) +
  facet_wrap(~ATM, ncol = 1, scales = 'free_y') +
  labs(title = 'ATM Forecasts for May 2010')

fc |> as_tsibble(index = DATE, key = ATM) |>
  autoplot(Cash) +
  autolayer(ATM, Cash, color = 'black') +
  facet_wrap(~ATM, ncol = 1, scales = 'free_y') +
  labs(title = 'ATM Usage')
```





Export to Excel

```
fc |>
  write.xlsx('./Output/ATMForecast.xlsx')
```

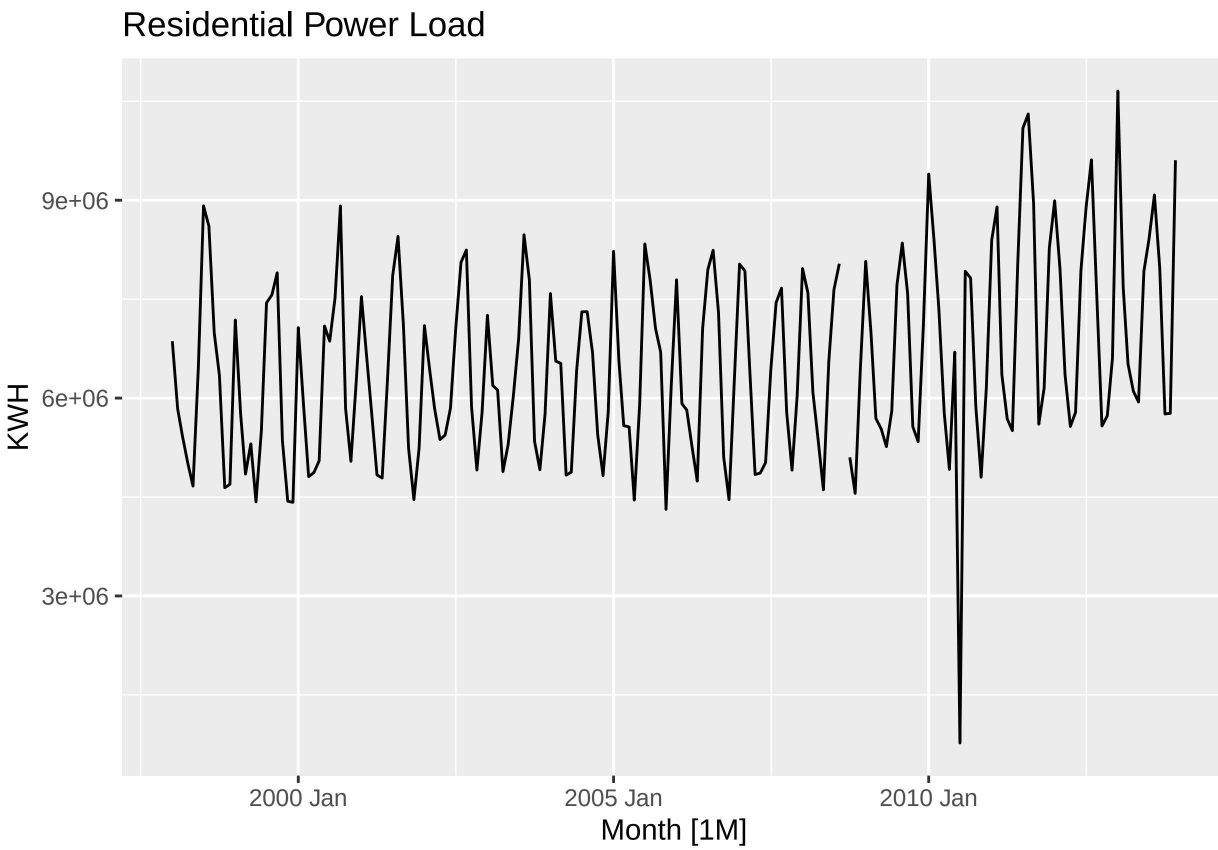
Part B

Forecasting Power - *ResidentialCustomerForecastLoad-624.xlsx*

Part B consists of a simple dataset of residential power usage for January 1998 until December 2013. Your assignment is to model these data and a monthly forecast for 2014. The data is given in a single file. The variable 'KWH' is power consumption in Kilowatt hours, the rest is straight forward. Add this to your existing files above.

Loading Data

Loaded data from an Excel file into a `tsibble` object. Converted the YYYY-MMM column to a yearmonth value, and excluded the CaseSequence column from the data.



Missing Values and Outliers

There is a single missing value and one value that appears to be a clear outlier.

```
Power |>
  mutate(mean = mean(KWH, na.rm = TRUE)) |>
  filter(is.na(KWH) | KWH < .25 * mean | KWH > 4 * mean)
```

```
## # A tsibble: 2 x 3 [1M]
##      Month    KWH     mean
##      <mth>  <dbl>    <dbl>
## 1 2008 Sep     NA 6502475.
## 2 2010 Jul 770523 6502475.
```

We will replace the KWH value for July 2010 with NA, then use an ARIMA model to interpolate the two values.

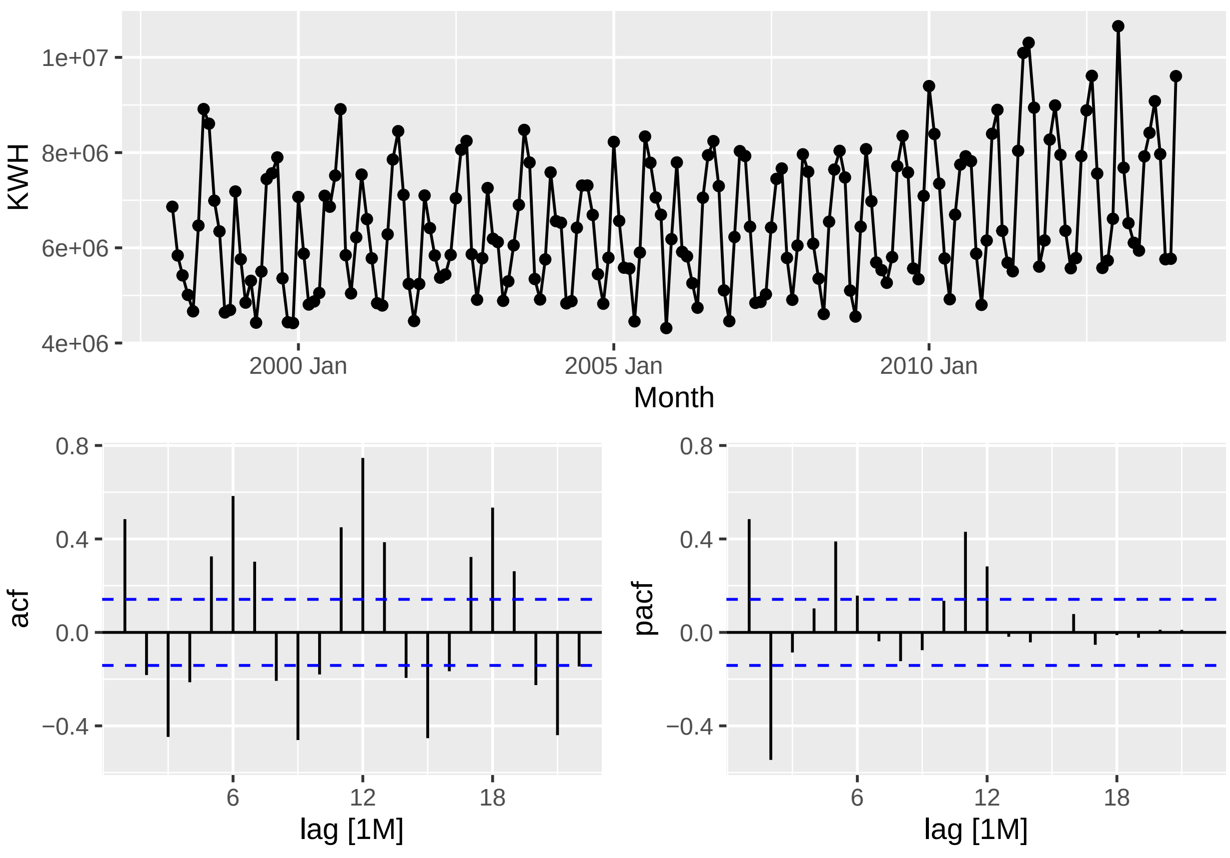
```
Power <- Power |>
  mutate(KWH = replace(KWH, Month == yearmonth('2010 Jul'), NA))

Power <- Power |>
  model(ARIMA(KWH)) |>
  interpolate(Power)

Power |>
  filter(Month == yearmonth('2008 Sep') | Month == yearmonth('2010 Jul'))
```

```
## # A tsibble: 2 x 2 [1M]
##      Month     KWH
##      <mth>   <dbl>
## 1 2008 Sep 7477566.
## 2 2010 Jul 7747778.
```

Series Exploration

The data has a clear seasonal pattern with spikes in winter and summer months and lows in the spring and fall. The ACF graph shows spikes on the lags in multiples of 6. There is an overall increasing trend in power consumption.

```
Power |>
  gg_tsdisplay(KWH, plot_type = 'partial')
```



Transformation

Identify a lambda value for a Box-Cox transformation.

```
lambda <- Power |>
  features(KWH, features = guerrero) |>
  pull(lambda_guerrero)

lambda
```

```
## [1] -0.2057366
```

Models

We will construct a few models to determine which is best.

```
Power.fit <- Power |>
  model(
    additive = ETS(box_cox(KWH, lambda) ~ error('A') + trend('N') + season('A')),
    multiplicative = ETS(box_cox(KWH, lambda) ~ error('M') + trend('N') + season('M')),
    arima = ARIMA(box_cox(KWH, lambda), stepwise = FALSE)
  )

Power.fit |>
  glance() |>
  select(.model:BIC) |>
  arrange(AIC)
```

```
## # A tibble: 3 x 6
##   .model              sigma2 log_lik    AIC   AICc    BIC
##   <chr>                <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
## 1 arima          0.0000141      759. -1504. -1503. -1481.
## 2 multiplicative 0.000000633    577. -1124. -1121. -1075.
## 3 additive       0.0000138      577. -1124. -1121. -1075.
```

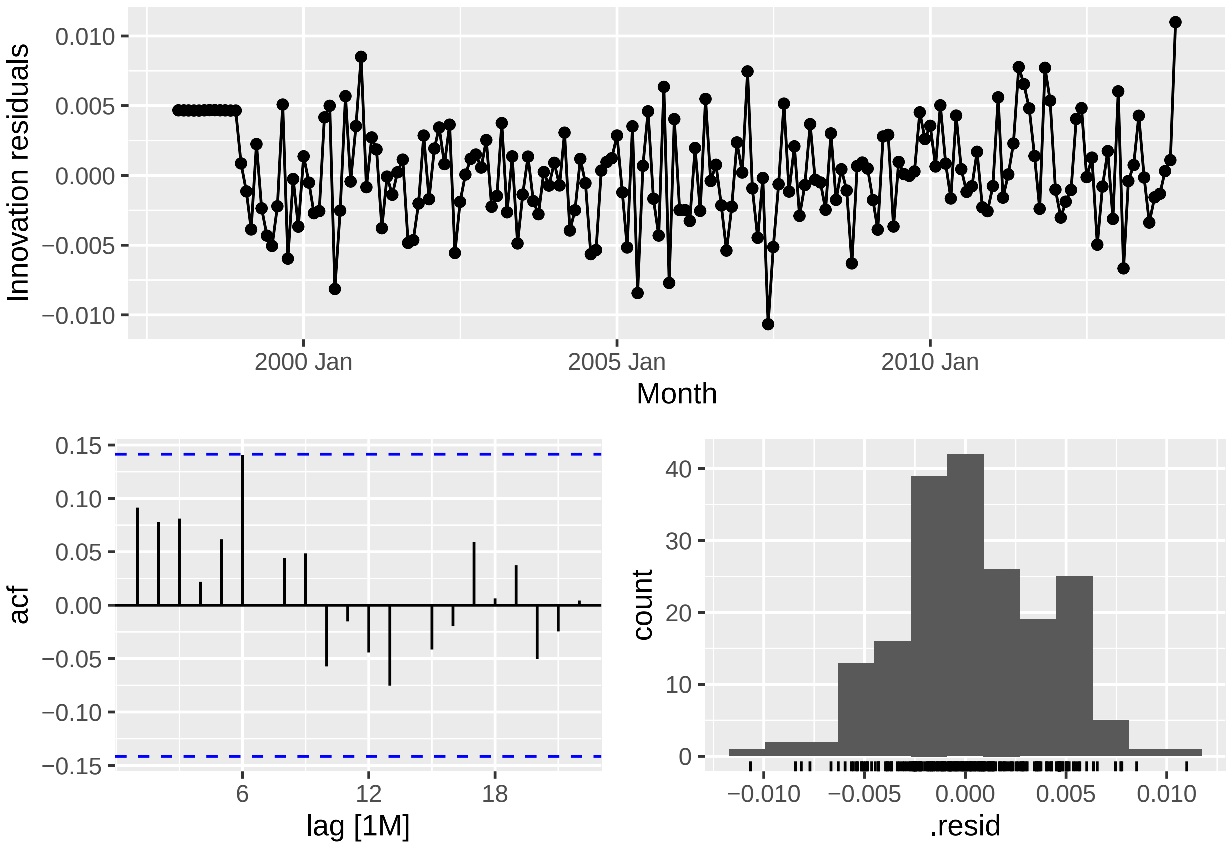
The ARIMA(0,0,3)(2,1,0)[12] model produced the lowest AIC.

```
Power.fit |>
  select(arima) |>
  report()
```

```
## Series: KWH
## Model: ARIMA(0,0,3)(2,1,0)[12] w/ drift
## Transformation: box_cox(KWH, lambda)
##
## Coefficients:
##          ma1     ma2     ma3     sar1     sar2  constant
##       0.2674  0.0620  0.2196  -0.7230  -0.3805    0.0013
## s.e.  0.0759  0.0849  0.0710   0.0745   0.0765    0.0005
##
## sigma^2 estimated as 1.414e-05:  log likelihood=758.83
## AIC=-1503.66   AICc=-1503.01   BIC=-1481.31
```

Graphing the residuals appear to be white noise. The ACF chart shows lags within the confidence interval. We conclude that the model is sound and can be used to forecast values.

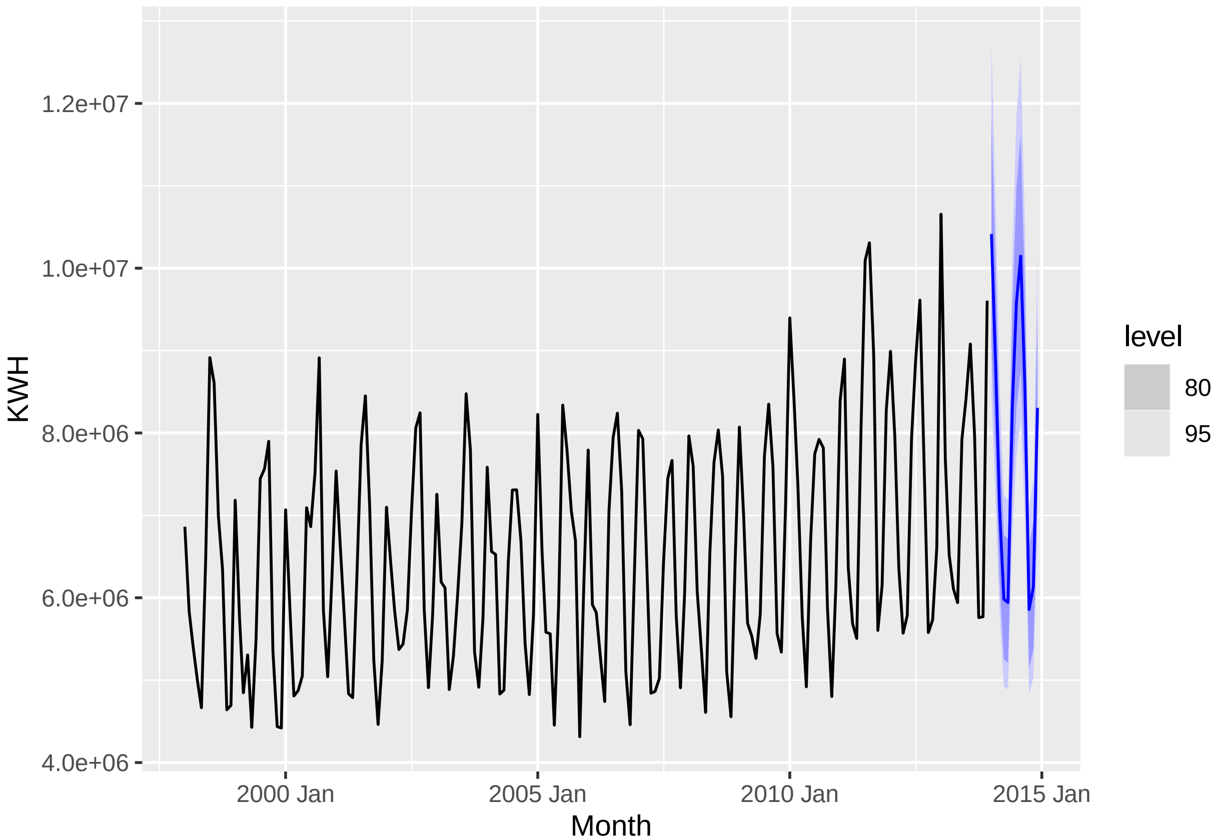
```
Power.fit |>
  select(arima) |>
  gg_tsresiduals()
```



Forecast

```
Power.fc <- Power.fit |>
  select(arima) |>
  forecast(h = 12)
```

```
Power.fc |>
  autoplot(Power)
```




```r
Power.fc |>
  as.data.frame() |>
  select(Month, .mean) |>
  rename(KWH = .mean) |>
  mutate(KWH = round(KWH,0))
```

```
##       Month      KWH
## 1  2014 Jan 10414868
## 2  2014 Feb  8780045
## 3  2014 Mar  7079855
## 4  2014 Apr  5982856
## 5  2014 May  5941894
## 6  2014 Jun  8305110
## 7  2014 Jul  9585835
## 8  2014 Aug 10146851
## 9  2014 Sep  8521497
## 10 2014 Oct  5856902
## 11 2014 Nov  6131467
## 12 2014 Dec  8304252
```

Export to Excel

```r
Power.fc |>
  as.data.frame() |>
```

```
  select(Month, .mean) |>
  rename(KWH = .mean) |>
  mutate(KWH = round(KWH,0)) |>
  write.xlsx('./Output/PowerForecast.xlsx')
```