# DATA 624: PREDICTIVE ANALYTICS Project 1

Gabriel Campos

Last edited March 23, 2024

```
library(fpp3)
library(dplyr)
library(ggplot2)
library(readxl)
library(tsibble)
library(psych)
library(tidyr)
library(forecast)
```

## Description

This project consists of 3 parts - two required and one bonus and is worth 15% of your grade. The project is due at 11:59 PM on Sunday Apr 11. I will accept late submissions with a penalty until the meetup after that when we review some projects.

### 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.

### 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.

### Part C

BONUS, optional (part or all), Waterflow_Pipe1.xlsx and Waterflow_Pipe2.xlsx

Part C consists of two data sets. These are simple 2 columns sets, however they have different time stamps. Your optional assignment is to time-base sequence the data and aggregate based on hour (example of what this looks like, follows). Note for multiple recordings within an hour, take the mean. Then to determine if the data is stationary and can it be forecast. If so, provide a week forward forecast and present results via Rpubs and .rmd and the forecast in an Excel readable file.

# EDA & Cleanup

## Data Load

https://github.com/GitableGabe/Data624_Data/raw/main/ATM624Data.xlsx

```
atm_coltype<-c("date","text","numeric")

atm_import<-read_xlsx('ATM624Data.xlsx', col_types = atm_coltype)
power_raw<-read_xlsx('ResidentialCustomerForecastLoad-624.xlsx')
# Ommitting Extra Credit as I won't be working on it
# WP1_df<-read_xlsx('Waterflow_Pipe1.xlsx')
# WP2_df<-read_xlsx('Waterflow_Pipe2.xlsx')
```

```
head(atm_import%>%
       filter(ATM=="ATM4"))
```

```
## # A tibble: 6 x 3
##   DATE                ATM    Cash
##   <dttm>              <chr> <dbl>
## 1 2009-05-01 00:00:00 ATM4  777.
## 2 2009-05-02 00:00:00 ATM4  524.
## 3 2009-05-03 00:00:00 ATM4  793.
## 4 2009-05-04 00:00:00 ATM4  908.
## 5 2009-05-05 00:00:00 ATM4   52.8
## 6 2009-05-06 00:00:00 ATM4   52.2
```

```
atm_range<-range(atm_import$DATE)
atm_range[1]
```

```
## [1] "2009-05-01 UTC"
```

```
atm_range[2]
```

```
## [1] "2010-05-14 UTC"
```

```
sapply(atm_import, function(x) sum(is.na(x)))
```

```
## DATE  ATM Cash
##    0   14   19
```

```
data.frame(atm_import$DATE[atm_import$Cash %in% NA])
```

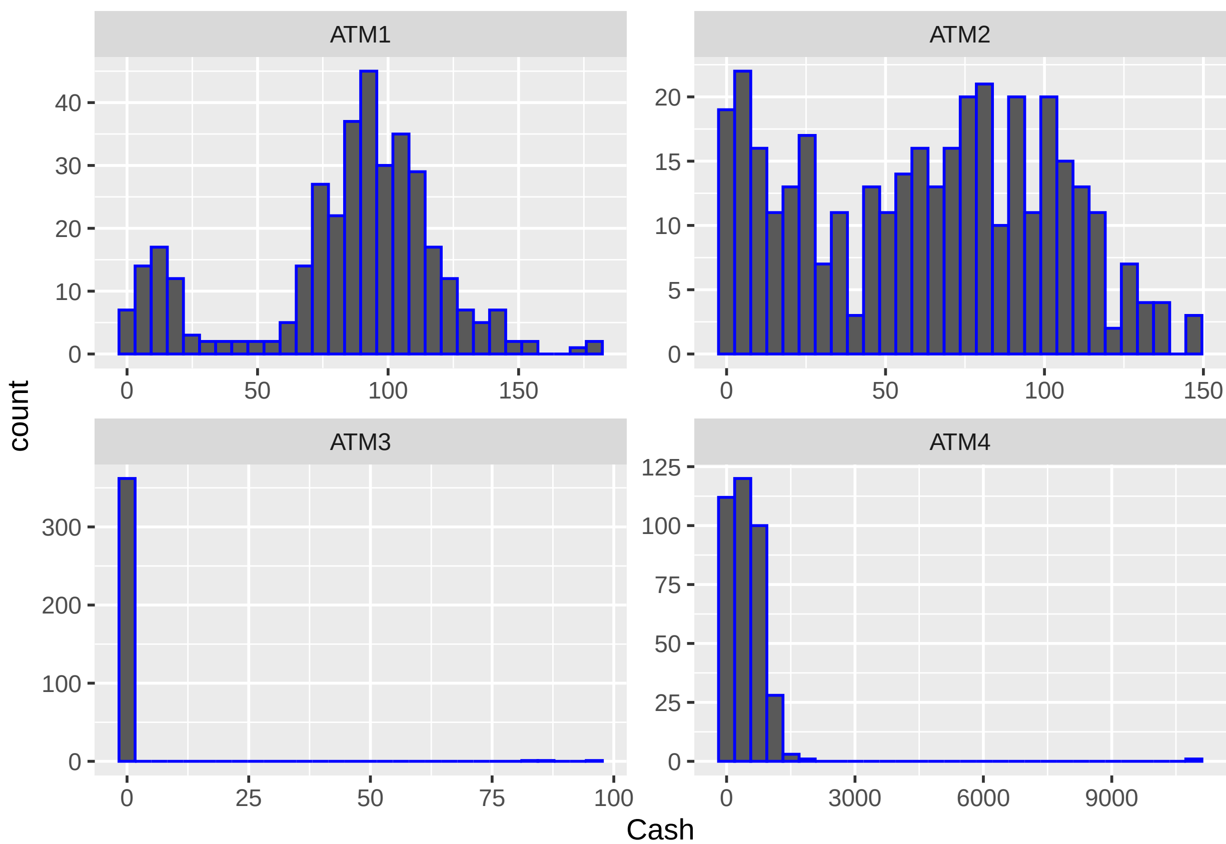
```
##    atm_import.DATE.atm_import.Cash..in..NA.
## 1                                2009-06-13
## 2                                2009-06-16
## 3                                2009-06-18
## 4                                2009-06-22
## 5                                2009-06-24
## 6                                2010-05-01
## 7                                2010-05-02
## 8                                2010-05-03
## 9                                2010-05-04
## 10                               2010-05-05
## 11                               2010-05-06
```

```
## 12                              2010-05-07
## 13                              2010-05-08
## 14                              2010-05-09
## 15                              2010-05-10
## 16                              2010-05-11
## 17                              2010-05-12
## 18                              2010-05-13
## 19                              2010-05-14
```

- ATM624Data had attribute type mismatches, and was converted on import.
- Date conversion somehow kept date time as POSIXct
- ATM4 shows values in greater decimals any country, with Dinars being the only Country that uses more than 2 decimals when using its currency, but even the dinar stops at the 100th decimal.
- Date range is 05-01-2009 to 05-14-2010
- we see the count of NAs in ATM is 14 and Cash column is 19
- The NA dates vary and are not exclusive to a specific sequential time period that we can just filter out.
- I am curious about the distrobution of cash considering the forecast ask for this project.

```
atm_import %>%
  filter(DATE < "2010-05-01", !is.na(ATM)) %>%
  ggplot(aes(x = Cash)) +
    geom_histogram(bins = 30, color= "blue") +
    facet_wrap(~ ATM, ncol = 2, scales = "free")
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```r
(atm_df <- atm_import %>%
  mutate(DATE = as.Date(DATE)) %>%
   filter(DATE<"2010-05-01")%>%
  pivot_wider(names_from=ATM, values_from = Cash))
```

```
## # A tibble: 365 x 5
##    DATE        ATM1  ATM2  ATM3  ATM4
##    <date>     <dbl> <dbl> <dbl> <dbl>
##  1 2009-05-01    96   107     0 777.
##  2 2009-05-02    82    89     0 524.
##  3 2009-05-03    85    90     0 793.
##  4 2009-05-04    90    55     0 908.
##  5 2009-05-05    99    79     0  52.8
##  6 2009-05-06    88    19     0  52.2
##  7 2009-05-07     8     2     0  55.5
##  8 2009-05-08   104   103     0 559.
##  9 2009-05-09    87   107     0 904.
## 10 2009-05-10    93   118     0 879.
## # i 355 more rows
```

```r
atm_df<-atm_df%>%
  as_tsibble(index=DATE)
head(atm_df)
```

```
## # A tsibble: 6 x 5 [1D]
##   DATE        ATM1  ATM2  ATM3  ATM4
##   <date>     <dbl> <dbl> <dbl> <dbl>
## 1 2009-05-01    96   107     0 777.
## 2 2009-05-02    82    89     0 524.
## 3 2009-05-03    85    90     0 793.
## 4 2009-05-04    90    55     0 908.
## 5 2009-05-05    99    79     0  52.8
## 6 2009-05-06    88    19     0  52.2
```

```r
summary(atm_df)
```

```
##       DATE                 ATM1             ATM2             ATM3
##  Min.   :2009-05-01   Min.   :  1.00   Min.   :  0.00   Min.   : 0.0000
##  1st Qu.:2009-07-31   1st Qu.: 73.00   1st Qu.: 25.50   1st Qu.: 0.0000
##  Median :2009-10-30   Median : 91.00   Median : 67.00   Median : 0.0000
##  Mean   :2009-10-30   Mean   : 83.89   Mean   : 62.58   Mean   : 0.7206
##  3rd Qu.:2010-01-29   3rd Qu.:108.00   3rd Qu.: 93.00   3rd Qu.: 0.0000
##  Max.   :2010-04-30   Max.   :180.00   Max.   :147.00   Max.   :96.0000
##                       NA's   :3        NA's   :2
##       ATM4
##  Min.   :    1.563
##  1st Qu.:  124.334
##  Median :  403.839
##  Mean   :  474.043
##  3rd Qu.:  704.507
##  Max.   :10919.762
##
```

```r
atm_df[!complete.cases(atm_df), ]
```

```
## # A tsibble: 5 x 5 [1D]
```

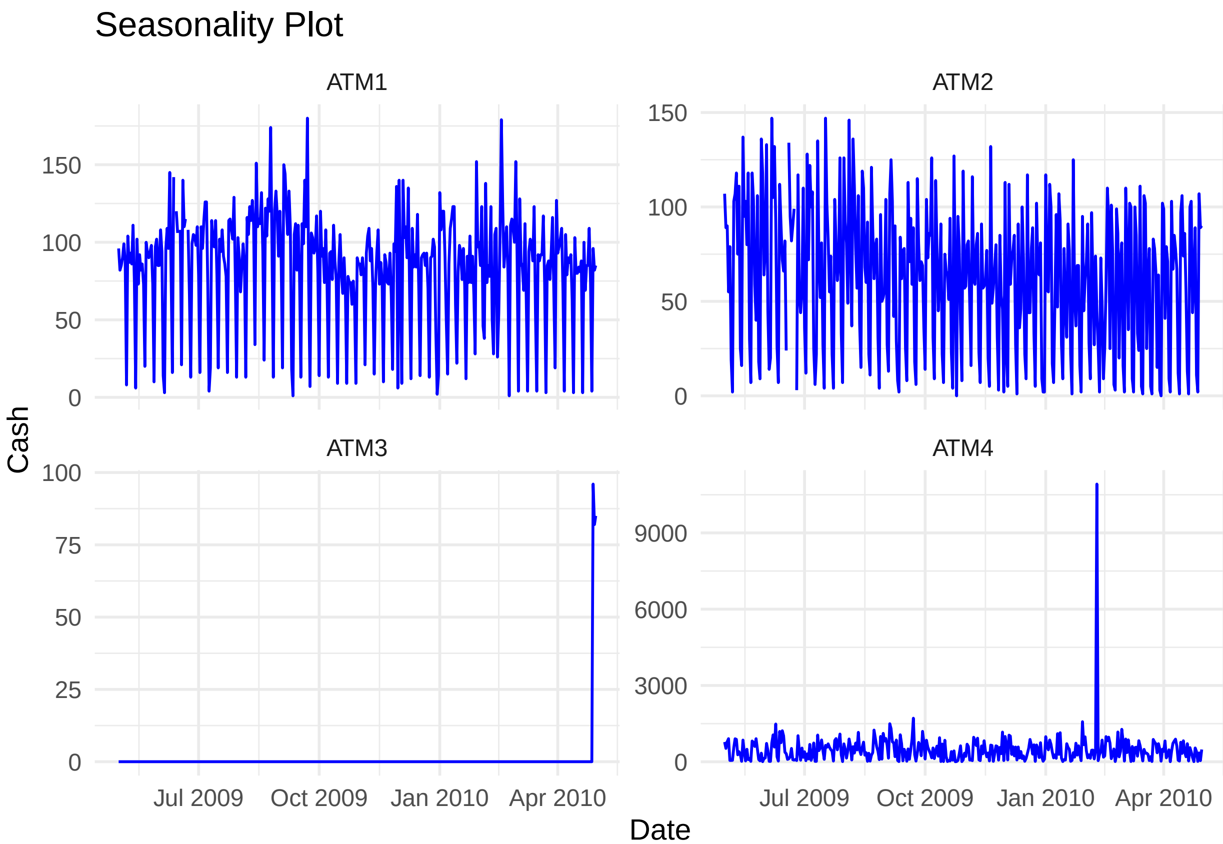
```
##   DATE        ATM1  ATM2  ATM3  ATM4
##   <date>     <dbl> <dbl> <dbl> <dbl>
## 1 2009-06-13    NA    91     0 746.
## 2 2009-06-16    NA    82     0 373.
## 3 2009-06-18    21    NA     0  92.5
## 4 2009-06-22    NA    90     0  80.6
## 5 2009-06-24    66    NA     0  90.6
```

```r
atm_df%>%
  select(DATE,ATM3)%>%
  filter(ATM3>0)
```

```
## # A tsibble: 3 x 2 [1D]
##   DATE        ATM3
##   <date>     <dbl>
## 1 2010-04-28    96
## 2 2010-04-29    82
## 3 2010-04-30    85
```

- Converting `DATE` into a date value made senses type POSIXct may cause future issues.
- Pivoting allowed us to separate the ATM's categorically and isolate the NAs for removal.
- We are able to see that five entries contain NAs and the dates all reside in June
- ATM3 only has 3 dates with withdrawals 4-28 through 4-30 or 2010
- These results also brings to question whether there may be some seasonality that will impact May's forecasting
- Considering the distrobution, I chose to replace the missing values with the median, as the skewed values in ATM 3 & 4 I believe with negatively impact the mean

```r
# seasonality
atm_import %>%
  filter(DATE < "2010-05-01", !is.na(ATM)) %>%
  ggplot(aes(x = DATE, y = Cash, col = ATM)) +
    geom_line(color="blue") +
    facet_wrap(~ ATM, ncol = 2, scales = "free_y")+
  labs(title = "Seasonality Plot", x = "Date", y = "Cash") +
    theme_minimal()
```




```r
median_value <- median(atm_df[["ATM1"]], na.rm = TRUE)
atm_df[["ATM1"]][is.na(atm_df[["ATM1"]])] <- median_value
median_value <- median(atm_df[["ATM2"]], na.rm = TRUE)
atm_df[["ATM2"]][is.na(atm_df[["ATM2"]])] <- median_value
```

```r
atm_df[!complete.cases(atm_df), ]
```

```
## # A tsibble: 0 x 5 [?]
## # i 5 variables: DATE <date>, ATM1 <dbl>, ATM2 <dbl>, ATM3 <dbl>, ATM4 <dbl>
```

# Forecasts

## ATM1

### STL Decomposition

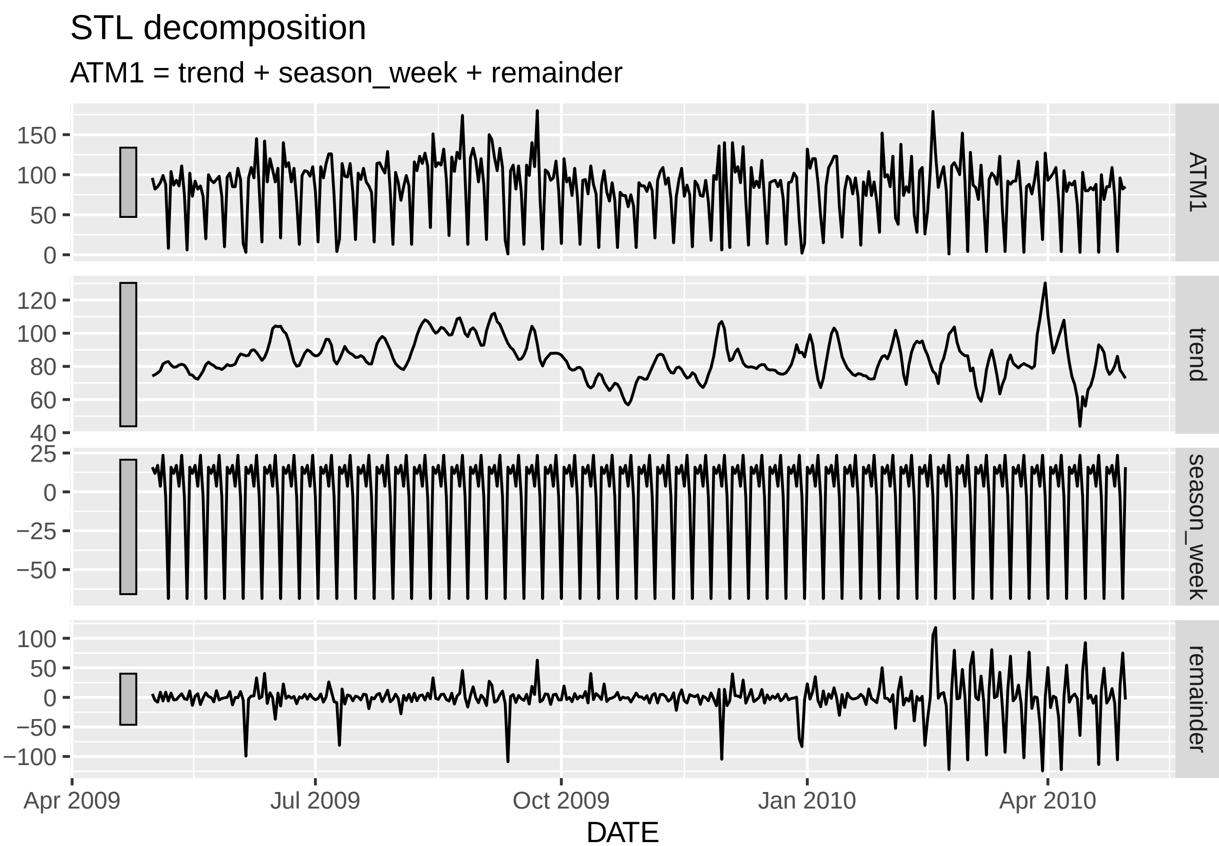
The seasonality plot did not show a trend in the long term but a better assessment in weekly interval is likely needed, using resources from Rob J Hyndman and George Athanasopoulos, Forecasting: Principles and Practice (3rd ed) section 3.6 STL decomposition I will perform a STL "Seasonal and Trend decomposition using Loess" decomposition of the series. To make it weekly I'll set the parameter `trend(window = 7)` and the `season(window='periodic')` to impose seasonality element across days of the week.

My reference come directly from the chapter.

```r
us_retail_employment |>
  model(
    STL(Employed ~ trend(window = 7) +
                   season(window = "periodic"),
    robust = TRUE)) |>
```

```
            components() |>
            autoplot()
```

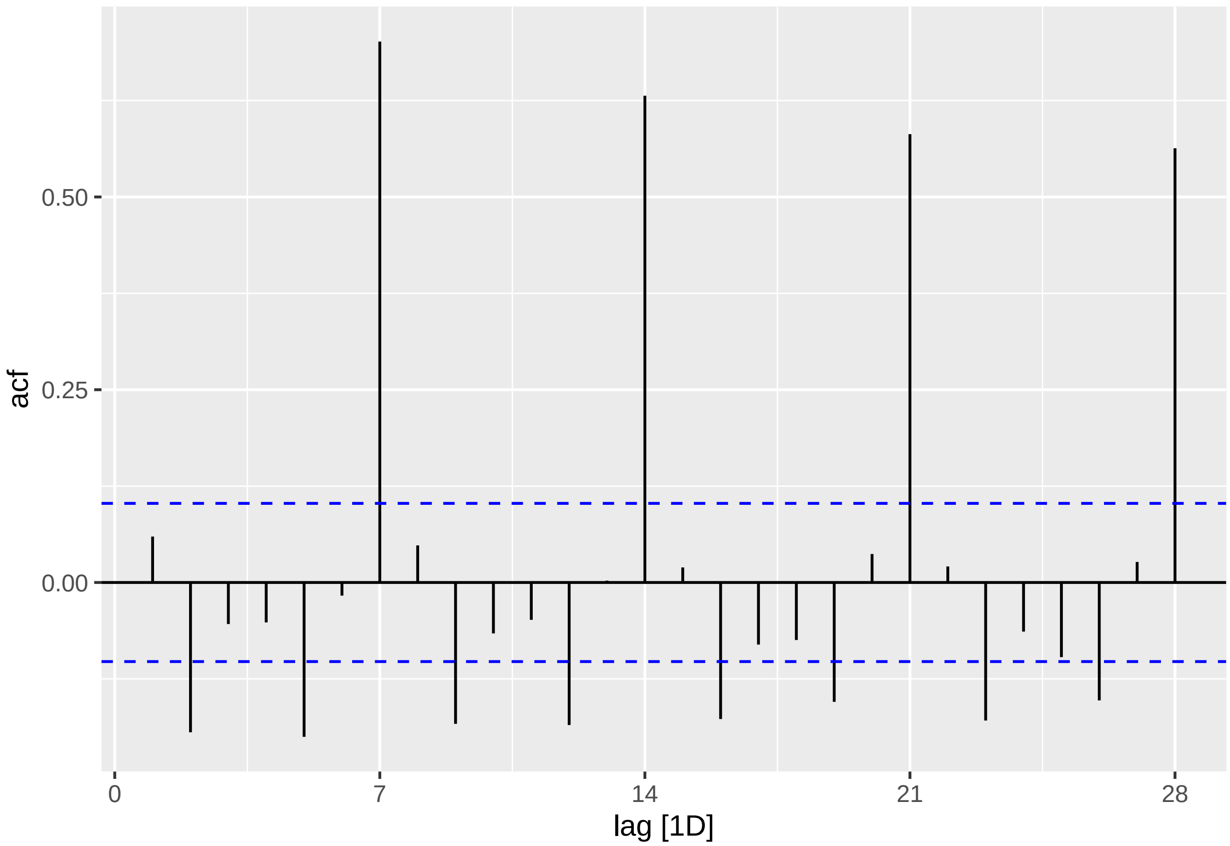
```
atm1_df <- atm_df %>%
  dplyr::select(DATE, ATM1)

atm1_df %>%
  model(
    STL(ATM1 ~ trend(window = 7) +
                season(window = "periodic"),
    robust = TRUE)) %>%
  components() %>%
  autoplot()
```




```
ndiffs(atm1_df$ATM1)
```

```
## [1] 0
```

```
atm1_df %>%
  ACF(ATM1, lag_max = 28) %>%
  autoplot()
```



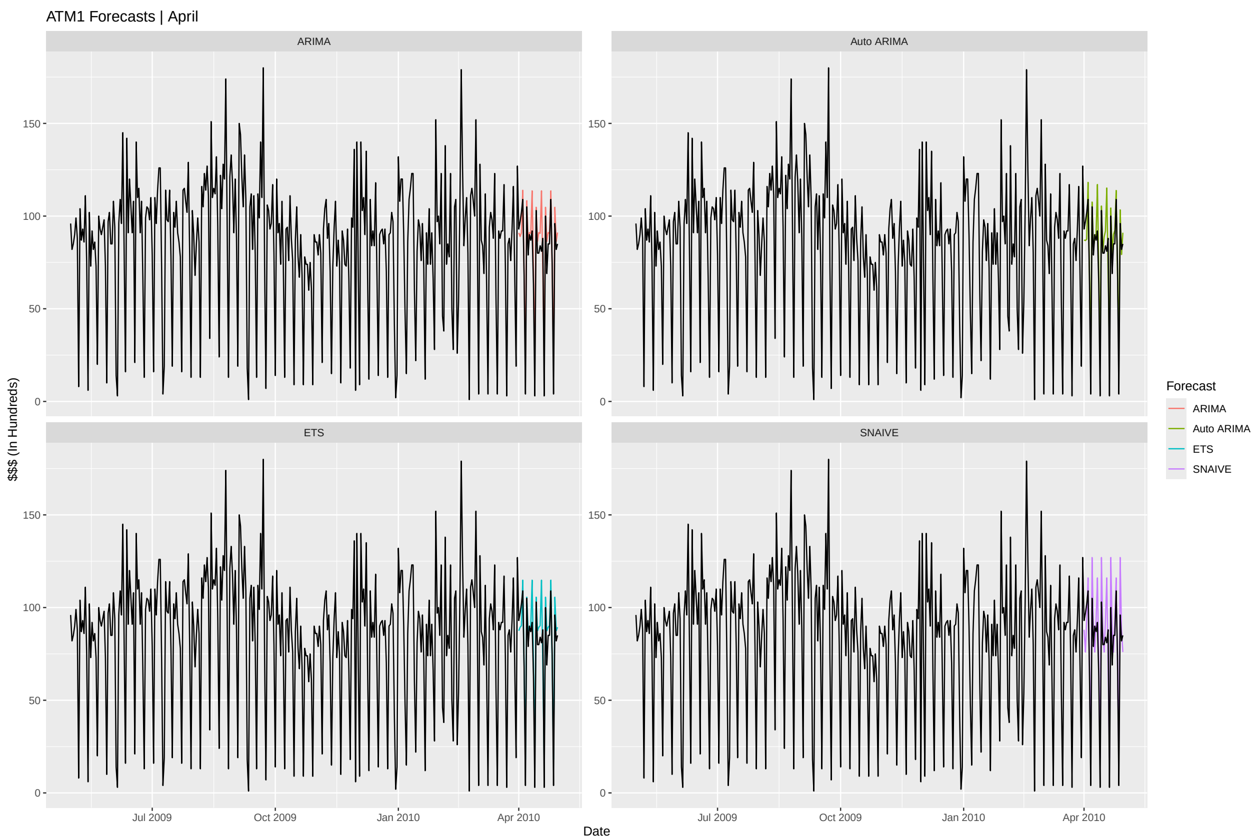

The STL decomposition wasn't as telling as I would have liked, however the ACF plot presents lags at 2, 5, and 7. I believe, given the week starts on Sunday, that this represents Monday, Thursday and Saturday as the days with the most lag. 7 has shown the value with the most significant lag. There is a decreasing trend with the ACF plot, and supports that the data is non-stationary would require differencing however $r_1's$ small value and the results of the `ndiff()` function, showing the first number of differences as 0, negates that suspicion.

### ARIMA

Seasonal naive method was my preferred choice considering the seasonality, and so we can use the prior time period's withdrawals to conduct our forecast, but I also like to default to `Auto ARIMA` for the optimized selection. I assume ETS and ARIMA wont perform as well but will await for the comparisons.

```
# train
train <- atm1_df %>%
  filter(DATE < "2010-04-01")


fit <- train %>%
  model(
    SNAIVE = SNAIVE(ATM1),
    ETS = ETS(ATM1),
    ARIMA = ARIMA(ATM1),
    `Auto ARIMA` = ARIMA(ATM1, stepwise = FALSE, approx = FALSE)
  )

# forecast April
forecast_ATM1 <- fit %>%
```

```r
  forecast(h = 30)

#plot
forecast_ATM1 %>%
  autoplot(atm1_df, level = NULL)+
  facet_wrap( ~ .model, scales = "free_y") +
  guides(colour = guide_legend(title = "Forecast"))+
  labs(title= "ATM1 Forecasts | April") +
  xlab("Date") +
  ylab("$$$ (In Hundreds)")
```




```r
# RMSE
accuracy(forecast_ATM1, atm1_df) %>%
  select(.model, RMSE:MAPE)
```

```
## # A tibble: 4 x 5
##   .model       RMSE   MAE    MPE  MAPE
##   <chr>       <dbl> <dbl>  <dbl> <dbl>
## 1 ARIMA        12.3  9.68  -82.9  86.0
## 2 Auto ARIMA   13.9 10.5  -103.  107.
## 3 ETS          11.6  9.13  -61.5  65.1
## 4 SNAIVE       16.0 13.5   -66.2  73.4
```

When interpreting the results, the model with the lowest RMSE and MAE value and the MPE and MAPE values closes to zero the best performing. This is true in all cases for ETS indicating it is the best performing.