DATA 624: Project 1

Juliann McEachern

October 22, 2019

Contents

Overview

I am leaving the project overview page here for us to compile our final report in one singular document. We will add additional information here regarding project one to include explanation of process, etc.

Dependencies

Please add all libraries used here.

The following R libraries were used to complete Project 1:

```
# General
library('easypackages')

libraries('knitr', 'kableExtra', 'default')

# Processing
libraries('readxl', 'tidyverse', 'janitor', 'lubridate')

# Graphing
libraries('ggplot2', 'grid', 'gridExtra', 'ggfortify','ggpubr')

# Timeseries
libraries('zoo', 'urca', 'tseries', 'timetk')

# Math
libraries('forecast')
```

Data

Data was stored within our group repository and imported below using the `readxl` package. Each individual question was solved within an R script and the data was sourced into our main report for discussion purposes. The R scripts are available within our appendix for replication purposes.

For grading purposes, we exported and saved all forecasts as a csv in our data folder.

```
# Data Aquisition
atm_data <- read_excel("data/ATM624Data.xlsx")
power_data <- read_excel("data/ResidentialCustomerForecastLoad-624.xlsx")
pipe1_data <- read_excel("data/Waterflow_Pipe1.xlsx")
pipe2_data <- read_excel("data/Waterflow_Pipe2.xlsx")

# Source Code
source("scripts/Part-A-JM.R")
```

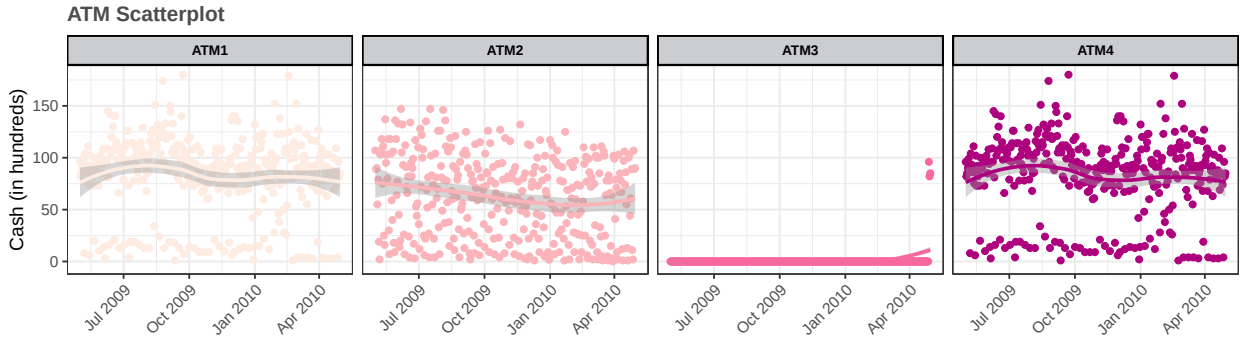
1 Part A

Instructions: 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. I am giving you data, please provide your written report on your findings, visuals, discussion and your R code all within a Word readable document, except the forecast which you will put in an Excel readable file. I must be able to cut and paste your R code and run it in R studio. Your report must be professional - most of all - readable, EASY to follow. Let me know what you are thinking, assumptions you are making! Your forecast is a simple CSV or Excel file that MATCHES the format of the data I provide.

1.1 Exploration

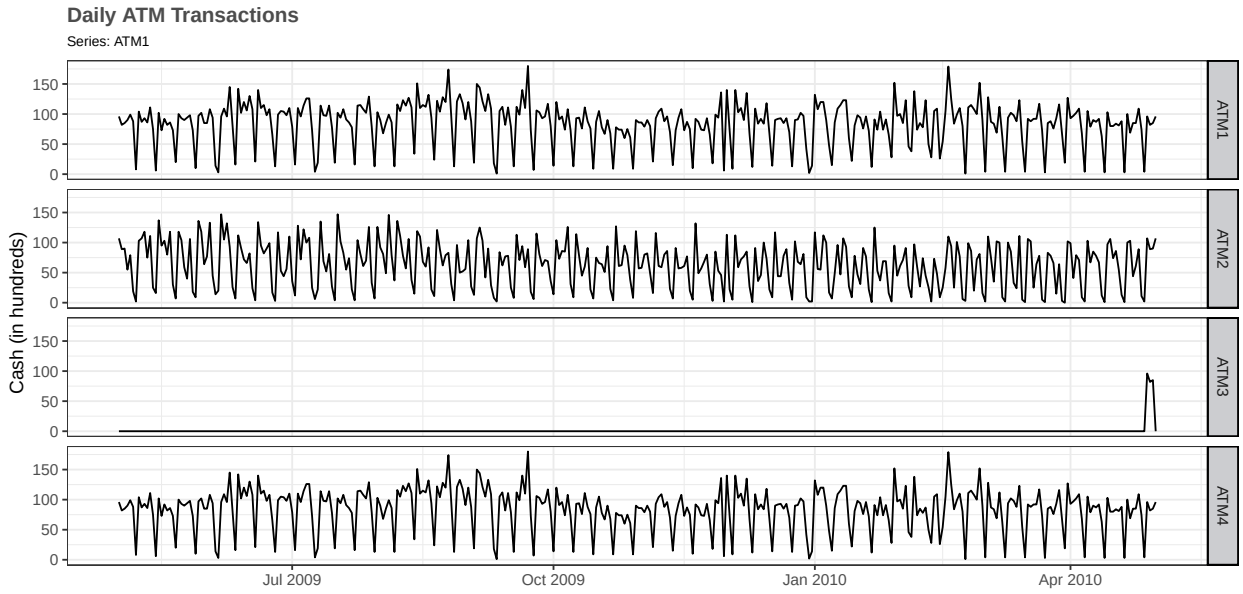
Through data exploration, we identified that the original data file contained `NA` values in our `ATM` and `Cash` columns for 14 observations in May 2010. We removed these missing values and transformed the dataset into a wide format. Our cleaned dataframe was then converted into a timeseries format using the `zoo` package for forecasting in the next section. Our initial review of the data showed that ATM2 contained one missing value on 2009-10-25 and that ATM4 contained a potential outlier of $1123 on 2010-02-09. We replaced both values with the corresponding mean value of each machine.

Next, we used a scatterplot to take an initial look at the correlation between cash withdrawals and dates for each machine. We can identified similiar patterns between ATM1 and ATM4, which show non-linear fluxuations that suggest a potential trend component in these timeseries. ATM2 follows a relatively linear path and decreases overtime. This changes in the last few observations, where withdrawals begin to increase. There are only 3 observed transactions for ATM3 that appear at the end of the captured time period.



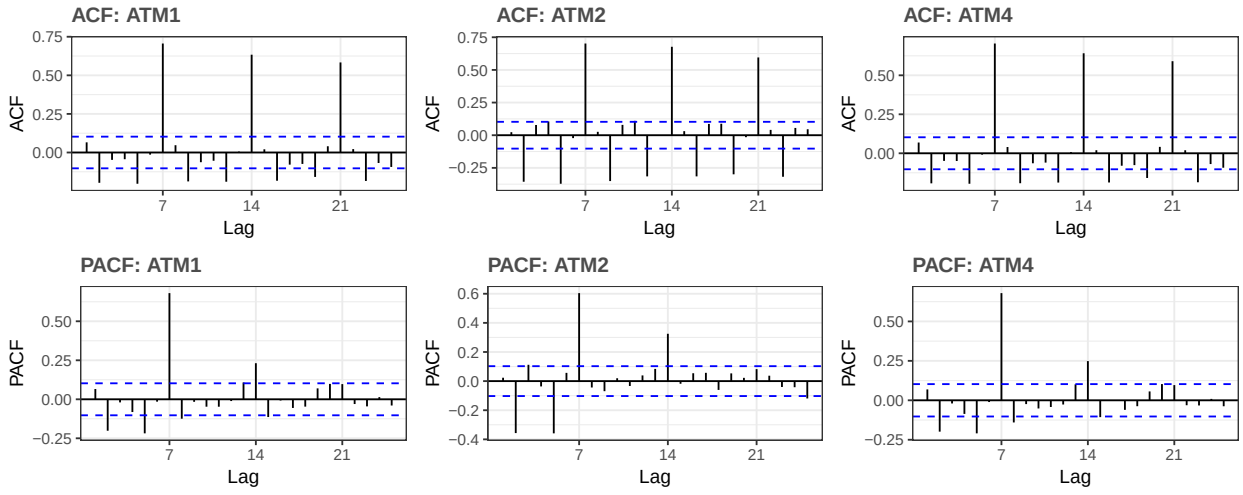
1.2 Timeseries Plots

As mentioned in our data exploration, the time series for ATM3 only contains 3 transactions, thus we deemed this series not suitable for modeling and forecasting. As a result, our following sections focus on evaluating, modeling, and forecasting transactions for only the ATM1, ATM2, and ATM4 series.



1.3 Evaluation

We constructed our timeseries using a weekly frequency. Our ACF plots for each ATM showcases large, decreasing lags starting at 7. This pattern continues in a multiple of seven, which confirms our assumption about seasonality within the observed data. These lags are indicative of a weekly pattern.



Our plots further suggest that the ATM data is non-stationary. We performed a unit root test using the `ur.kpss()` function to confirm this observation. The test results below show that second differencing is required on all three series.

Table 1.1: KPSS unit root test

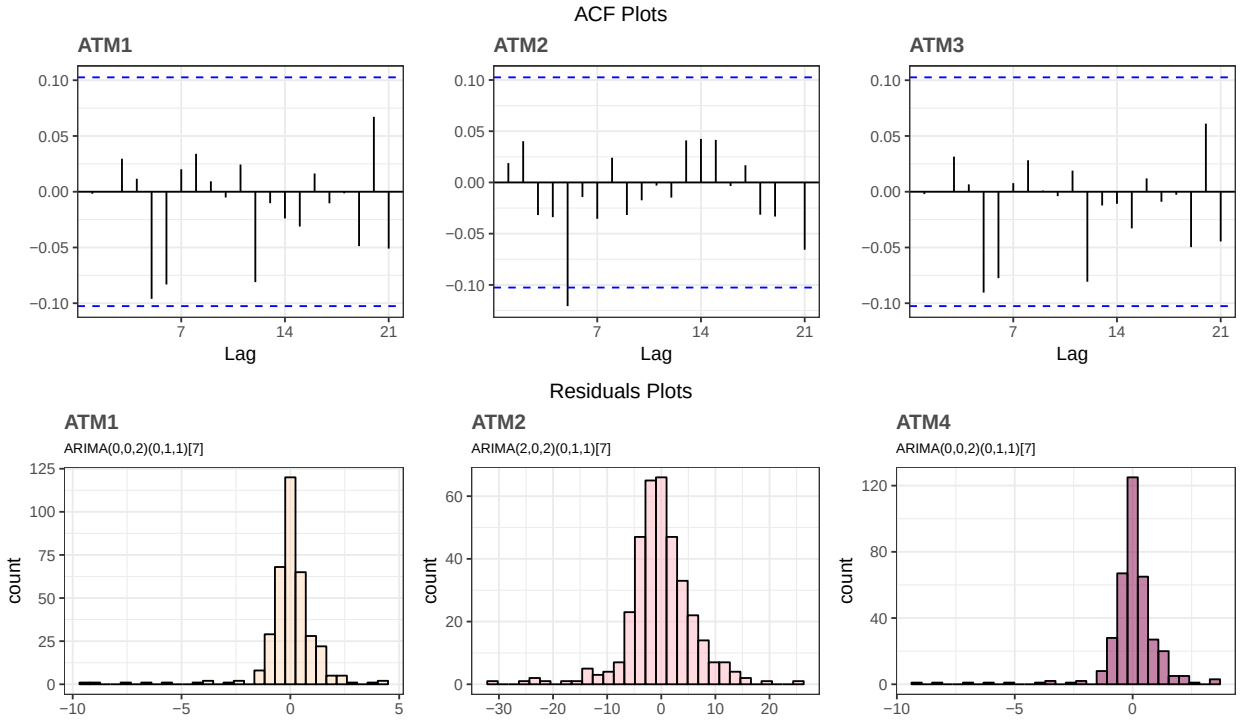
ATM	No-Diff	Diff-1
ATM1	0.4967	0.0219
ATM2	2.0006	0.016
ATM4	0.5182	0.0211

1.3.1 Modeling

We used `auto.arima()` and set `D=1` to account for seasonal differencing of our data to select the best ARIMA models. The full models and accuracy statistics for each series can be viewed in the appendix.

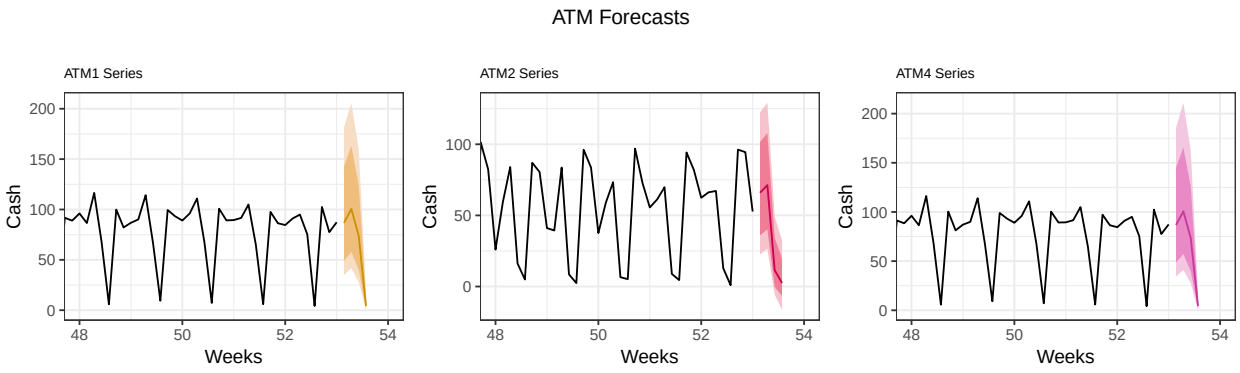
- **ATM1**: ARIMA$(0,0,2)(0,1,1)_7$
- **ATM2**: ARIMA$(2,0,2)(0,1,1)_7$
- **ATM4**: ARIMA$(0,0,2)(0,1,1)_7$

The following ACF plots show us that our differentiated data is now stationary. Further, the residual histograms follow a relatively normal distribution, which confirms that the models adequately fits the observed data.



1.4 Forecast

Finally, we applied a forecast to each series for 4 weeks of May 2010. The full forecasts can be viewed in the appendix section and are also located within our data output folder.



Appendix

Part A

ARIMA Model Summary

ATM1:

```
FALSE Series: ATM1_ts
FALSE ARIMA(0,0,2)(0,1,1)[7]
FALSE Box Cox transformation: lambda= 0.2584338
FALSE
FALSE Coefficients:
FALSE          ma1      ma2     sma1
FALSE       0.1085  -0.1089  -0.6425
FALSE s.e.  0.0524   0.0521   0.0431
FALSE
FALSE sigma^2 estimated as 1.726:  log likelihood=-606.1
FALSE AIC=1220.2   AICc=1220.32   BIC=1235.72
```

ATM2:

```
FALSE Series: ATM2_ts
FALSE ARIMA(2,0,2)(0,1,1)[7]
FALSE Box Cox transformation: lambda= 0.661752
FALSE
FALSE Coefficients:
FALSE           ar1      ar2     ma1     ma2     sma1
FALSE       -0.4238  -0.8978  0.4766  0.7875  -0.7064
FALSE s.e.   0.0592   0.0473  0.0883  0.0608   0.0417
FALSE
FALSE sigma^2 estimated as 38.94:  log likelihood=-1162.96
FALSE AIC=2337.93   AICc=2338.17   BIC=2361.21
```

ATM4:

```
FALSE Series: ATM4_ts
FALSE ARIMA(0,0,2)(0,1,1)[7]
FALSE Box Cox transformation: lambda= 0.2328582
FALSE
FALSE Coefficients:
FALSE          ma1      ma2     sma1
FALSE       0.1095  -0.1088  -0.6474
FALSE s.e.  0.0524   0.0523   0.0420
FALSE
FALSE sigma^2 estimated as 1.439:  log likelihood=-573.5
FALSE AIC=1154.99   AICc=1155.11   BIC=1170.52
```

Forecast Tables

Table 1.2: ATM1 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
53.14286	86.682223	48.937327	142.63088	34.8072635	181.28052
53.28571	100.569238	57.906034	163.01809	41.7309116	205.83358
53.42857	73.710292	40.220763	124.43478	27.9573875	159.92833
53.57143	4.229029	1.044442	11.78734	0.3790053	18.47789

Table 1.3: ATM2 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
53.14286	65.913008	35.8986677	101.54660	22.681696	122.41793
53.28571	71.267875	40.2481208	107.77822	26.412979	129.07899
53.42857	11.469466	-0.1794763	34.28206	-5.665135	49.36418
53.57143	2.464152	-6.7261243	19.59437	-16.373937	32.40892

Table 1.4: ATM4 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
53.14286	86.714879	48.483322	144.69401	34.4084146	185.45461
53.28571	100.581689	57.229595	165.56192	41.0835082	210.92017
53.42857	73.645362	39.882279	125.94463	27.7136726	163.18586
53.57143	4.221433	1.165285	11.35003	0.4840215	17.69213

R Script

```
# load data
atm_data <- read_excel("data/ATM624Data.xlsx")

# clean dataframe
atm <- atm_data %>%
  # create wide dataframe
  spread(ATM, Cash) %>%
  # remove NA column using function from janitor package
  remove_empty(which = "cols") %>%
  # filter unobserved values from May 2010
  filter(DATE < as.Date("2010-05-01")) %>%
  # ensure dates are ascending
  arrange(DATE)

## remove NA
atm$ATM2[is.na(atm$ATM2)] <- mean(atm$ATM2, na.rm = TRUE)

## remove outlier
atm$ATM4[which.max(atm$ATM4)] <- mean(atm$ATM4, na.rm = TRUE)

# create zoo time series
atm_zoo <- atm %>%
```

```r
  # remove column & generate date in timeseries using zoo
  select(-DATE) %>%
  # generate ts using zoo
  zoo(seq(from = as.Date("2009-05-01"), to = as.Date("2010-05-01"), by = 1))

# create standard time series
atm_ts <- atm %>%
  # remove column & generate date in timeseries using zoo
  select(-DATE) %>%
  # generate ts using zoo
  ts(start=1,  frequency = 7)

#subset data
ATM1_zoo <- atm_zoo[,1]; ATM1_ts <- atm_ts[,1]
ATM4_zoo <- atm_zoo[,4]; ATM4_ts <- atm_ts[,4]
ATM2_zoo <- atm_zoo[,2]; ATM2_ts <- atm_ts[,2]

#unit root test
## no diff
ATM1_ur <-ur.kpss(ATM1_ts)
ATM2_ur <-ur.kpss(ATM2_ts)
ATM4_ur <-ur.kpss(ATM4_ts)
## first order diff
ATM1d_ur <-ur.kpss(diff(ATM1_ts, lag=7))
ATM2d_ur <-ur.kpss(diff(ATM2_ts, lag=7))
ATM4d_ur <-ur.kpss(diff(ATM4_ts, lag=7))

# Modeling
## Lambda for Box-cox transformation
ATM1l <- BoxCox.lambda(ATM1_ts)
ATM2l <- BoxCox.lambda(ATM2_ts)
ATM4l <- BoxCox.lambda(ATM4_ts)

## ARIMA
ATM1_arima <-auto.arima(ATM1_ts, D = 1, lambda = ATM1l, approximation = F, stepwise = T)
ATM2_arima<-auto.arima(ATM2_ts, D = 1, lambda = ATM2l, approximation = F, stepwise = T)
ATM4_arima<-auto.arima(ATM4_ts, D = 1, lambda = ATM4l, approximation = F, stepwise = T)

# Forecast
ATM1_fc <- forecast(ATM1_arima,h=4)
ATM2_fc <- forecast(ATM2_arima,h=4)
ATM4_fc <- forecast(ATM4_arima,h=4)

# Save output
write.csv(ATM1_fc, file="forecasts/ATM1_Forecast.csv")
write.csv(ATM2_fc, file="forecasts/ATM2_Forecast.csv")
write.csv(ATM4_fc, file="forecasts/ATM4_Forecast.csv")
```