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')

# 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")
```

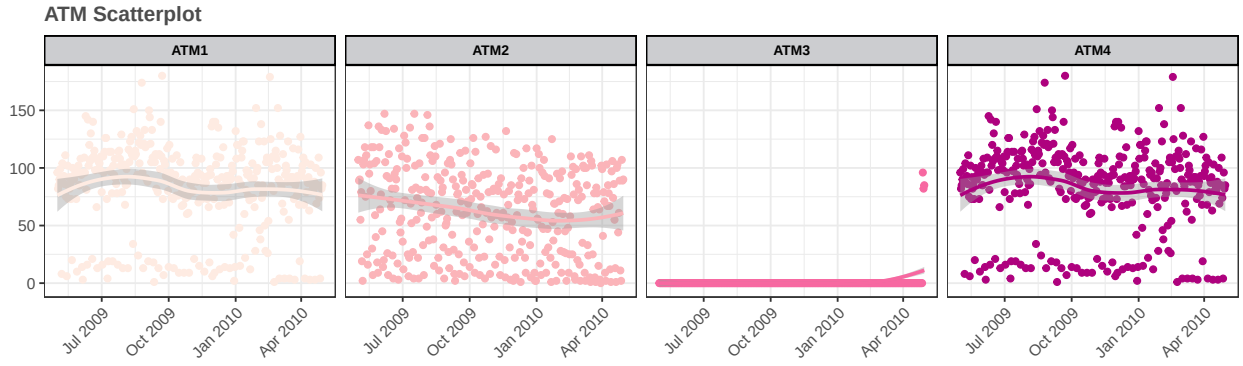
1 Part A

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

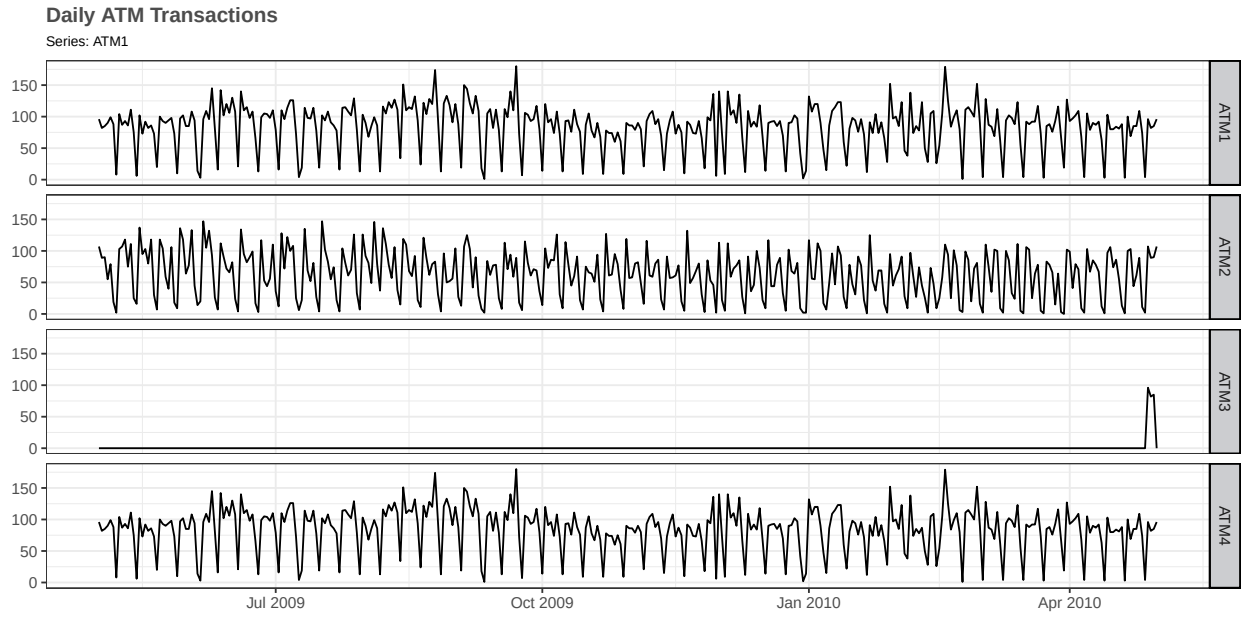
Through data exploration, we identified that the original data file contained `NA` values in our `ATM` and `Cash` columns for 14 observations in May 2010. We removed these missing values and transformed the dataset into a wide format. Our cleaned dataframe was then converted into a timeseries format 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.



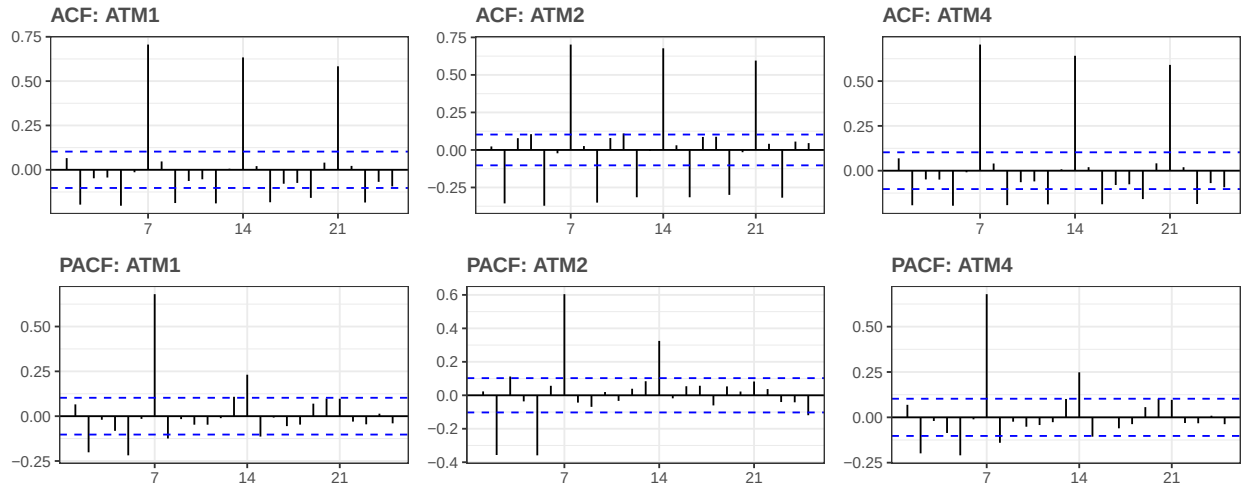
1.2 Timeseries Plots

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

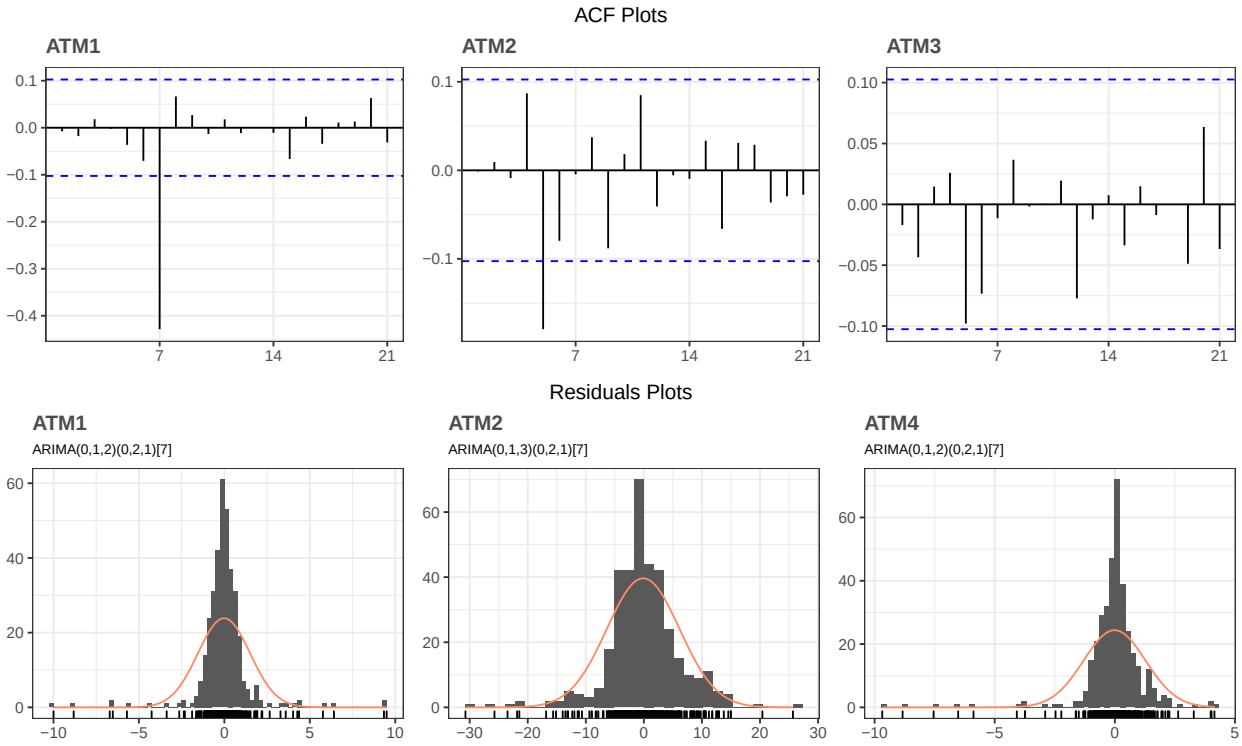
ATM	No-Diff	Diff-1	Log-Diff-1	Log-Diff-2
ATM1	0.4967	0.0219	0.0129	0.0077
ATM2	2.0006	0.016	NaN	NaN
ATM4	0.5182	0.0211	0.0128	0.0077

1.3.1 Modeling

We used `auto.arima()` on our differenced data to select the best ARIMA model for our series. The following models were selected based on AICc values:

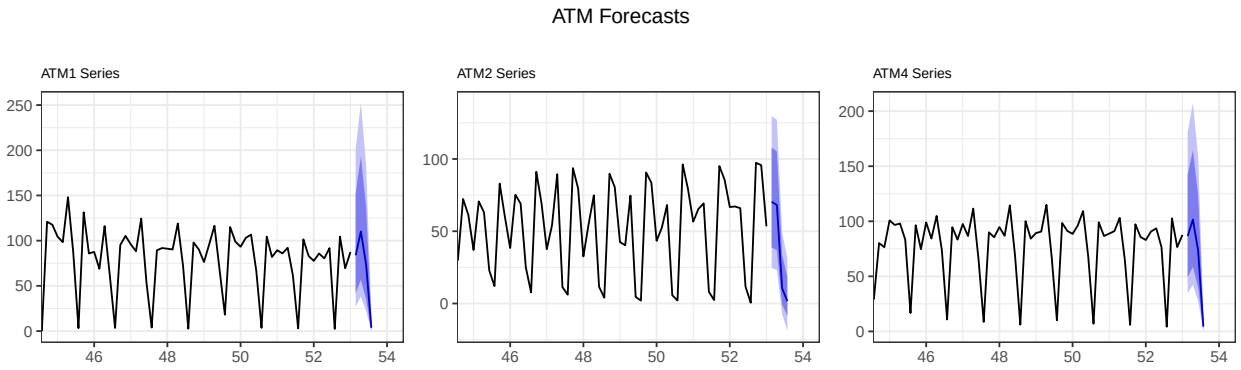
- **ATM1**: ARIMA(0,1,2)(0,2,1)$_7$ with errors
- **ATM2**: ARIMA(0,1,3)(0,2,1)$_7$ with errors
- **ATM4**: ARIMA(0,1,2)(0,2,1)$_7$ with errors

The following ACF plots show us that our differentiated data is now stationary and the residual histograms confirm that the model adequately fits the observed data.



1.3.2 Forecast

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



Appendix

Part A

Forecast Tables

Table 1.2: ATM1 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
53.14286	83.817966	41.4903974	151.93962	26.9523824	201.31621
53.28571	110.038431	57.0176652	192.99088	38.2480730	252.14714
53.42857	73.573807	35.1840074	136.67614	22.2984834	182.96563
53.57143	3.627256	0.5780783	12.53861	0.1288479	21.18631

Table 1.3: ATM2 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
53.14286	70.399901	38.764054	107.86146	24.778907	129.77780
53.28571	68.085302	36.835282	105.24426	23.106848	127.02440
53.42857	10.504707	-0.641800	33.57068	-7.075581	48.97866
53.57143	1.612731	-8.480086	18.35305	-19.006485	31.29272

Table 1.4: ATM4 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
53.14286	86.584988	48.985849	142.24258	34.8911333	180.66018
53.28571	101.530138	58.590610	164.29318	42.2872278	207.28680
53.42857	73.317040	40.109597	123.53235	27.9290828	158.63577
53.57143	4.356301	1.098715	12.00592	0.4070117	18.74379

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)
```

```r
## 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 %>%
  # 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))
## seasonal diff
ATM1sd_ur <-ur.kpss(diff(log(ATM1_ts), lag=7))
ATM2sd_ur <-ur.kpss(diff(log(ATM2_ts), lag=7))
ATM4sd_ur <-ur.kpss(diff(log(ATM4_ts), lag=7))
## seasonal diff-diff
ATM1sdd_ur <-ur.kpss(diff(diff(log(ATM1_ts)), lag=7))
ATM2sdd_ur <-ur.kpss(diff(diff(log(ATM2_ts)), lag=7))
ATM4sdd_ur <-ur.kpss(diff(diff(log(ATM4_ts)), lag=7))

# Modeling
## Lambda for Box-cox transformation
ATM1_lambda <- BoxCox.lambda(ATM1_ts)
ATM2_lambda <- BoxCox.lambda(ATM2_ts)
ATM4_lambda <- BoxCox.lambda(ATM4_ts)

## ARIMA
ATM1_arima<-Arima(ATM1_ts, order = c(1, 0, 1),
                  seasonal=list(order=c(0, 2, 1),period = 7),
                  lambda=ATM1_lambda,  method="ML")
```

```r
ATM2_arima<-Arima(ATM2_ts, order = c(2, 1, 2),
                  seasonal=list(order=c(2, 1, 2),period = 7),
                  lambda=ATM2_lambda,  method="ML")

ATM4_arima<-Arima(ATM4_ts, order = c(1, 0, 1),
                  seasonal=list(order=c(1, 1, 1),period = 7),
                  lambda=ATM1_lambda, method="ML")

# 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")
```