Project One

Group Two

Vinicio Haro
Sang Yoon (Andy) Hwang
Julian McEachern
Jeremy O'Brien
Bethany Poulin

22 October 2019

Contents

Overview

Project 1 Overview. Explanation of process, etc.

Dependencies

The following R libraries were used to complete Project 1.

```
# Processing

library(readxl)
library(tidyverse)
library(zoo)
library(janitor)

## Insert Additional Packages Here

# Graphing
library(ggplot2)
library(grid)
library(gridExtra)

## Insert Additional Packages Here

# Timeseries
library(zoo)

## Insert Additional Packages Here

# Math
library(forecast)
library(urca)

## Insert Additional Packages Here

# Formatting
require(knitr)
require(kableExtra)
require(default)
```

Data Aquisition

Data was stored within our group repository and imported below using the `readxl` package.

```
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")
```

Part A: Forecasting ATM Withdrawals (JM)

Juliann's Answer to #1.

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.

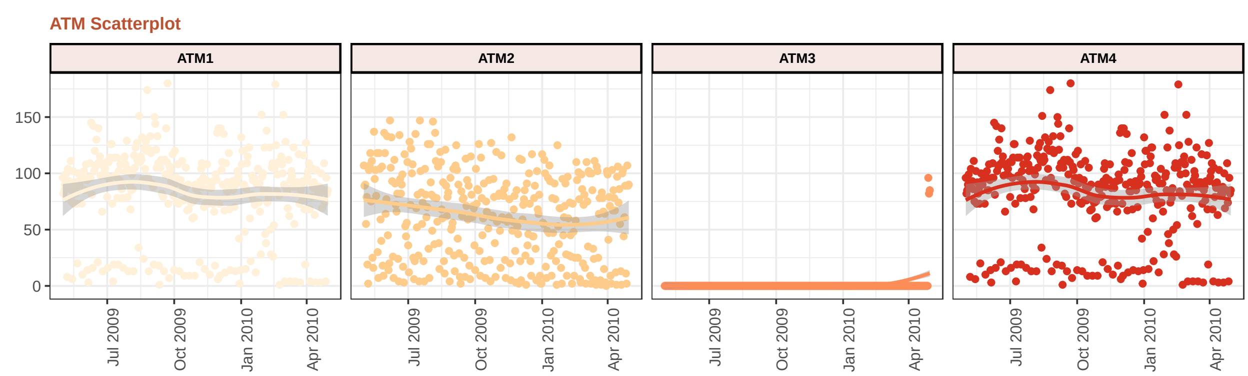
Exploration

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.

```r
# 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 time series
atm_ts <- 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("2010-05-14"), by = 1))

# plot atms as scatterplot
atm %>%
  # re-gather observations for facet plot
  gather(key=ATM, value=Cash, ATM1,ATM2, ATM3,ATM4) %>%
  # remove NA value from ATM2
  filter(complete.cases(.)) %>%
  # plot
  ggplot(aes(DATE, Cash, color=ATM)) +
  geom_point() +
  geom_smooth(method="loess") +
  facet_wrap(~ATM, scales='free_x', nrow=1) +
  labs(title="ATM Scatterplot")+
  theme_bw()+
  theme(legend.position = 'none')+
  scale_color_brewer()
```



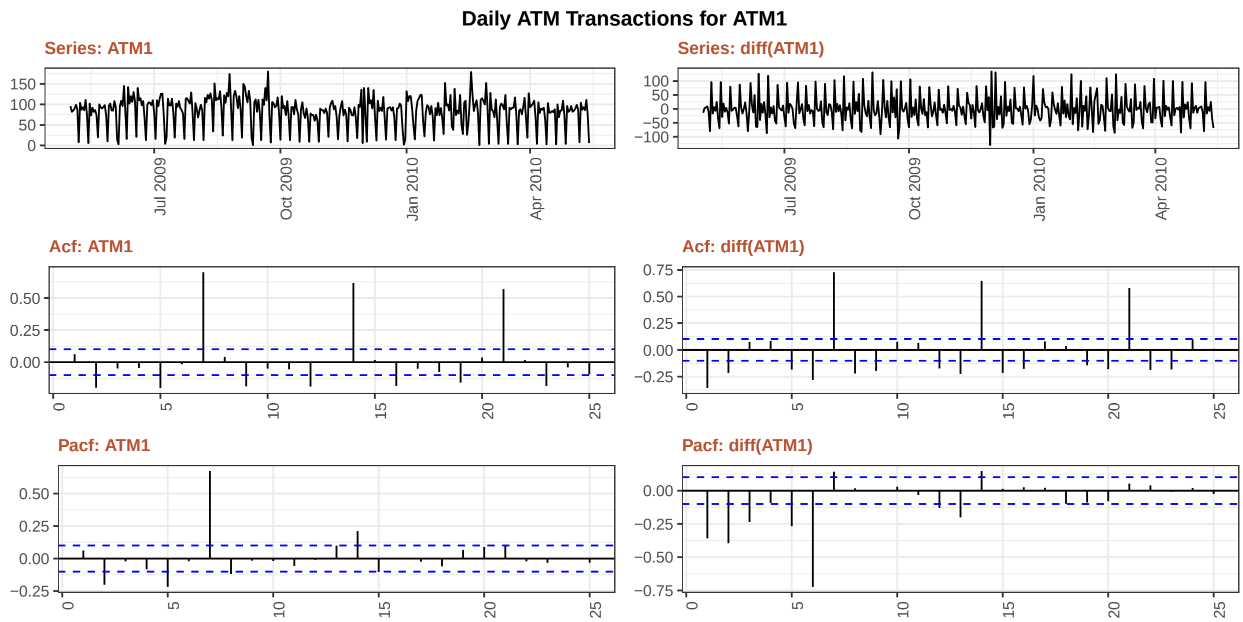
Forecast

We subsetted each atm series to apply unique forecasting methods based on the observed data.

ATM1

```
#subset data
ATM1 <- atm_ts[,1]

#differentiated
ATM1d <-  diff(ATM1)

p1<-autoplot(ATM1)+ labs(title="Series: ATM1")+
  theme_bw()+theme()
p2<-ggAcf(ATM1)+ labs(title="Acf: ATM1")+
  theme_bw()+theme()
```

```
p3<-ggPacf(ATM1)+ labs(title="Pacf: ATM1")+
  theme_bw()+theme()
p4<-autoplot(ATM1d)+ labs(title="Series: diff(ATM1)")+
  theme_bw()+theme()
p5<-ggAcf(ATM1d)+ labs(title="Acf: diff(ATM1)")+
  theme_bw()+theme()
p6<-ggPacf(ATM1d)+ labs(title="Pacf: diff(ATM1)")+
  theme_bw()+theme()

grid.arrange(grob=p1, p4, p2, p5, p3, p6,
             ncol=2,
             top=textGrob("Daily ATM Transactions for ATM1",
                          gp = gpar(fontface = "bold", cex = 1)))
```




```
# root test using urca package
ATM1 %>% diff() %>% ur.kpss() %>% summary()
```

```
FALSE
FALSE #######################
FALSE # KPSS Unit Root Test #
FALSE #######################
FALSE
FALSE Test is of type: mu with 5 lags.
FALSE
FALSE Value of test-statistic is: 0.0168
FALSE
FALSE Critical value for a significance level of:
FALSE                 10pct  5pct 2.5pct  1pct
FALSE critical values 0.347 0.463  0.574 0.739
```

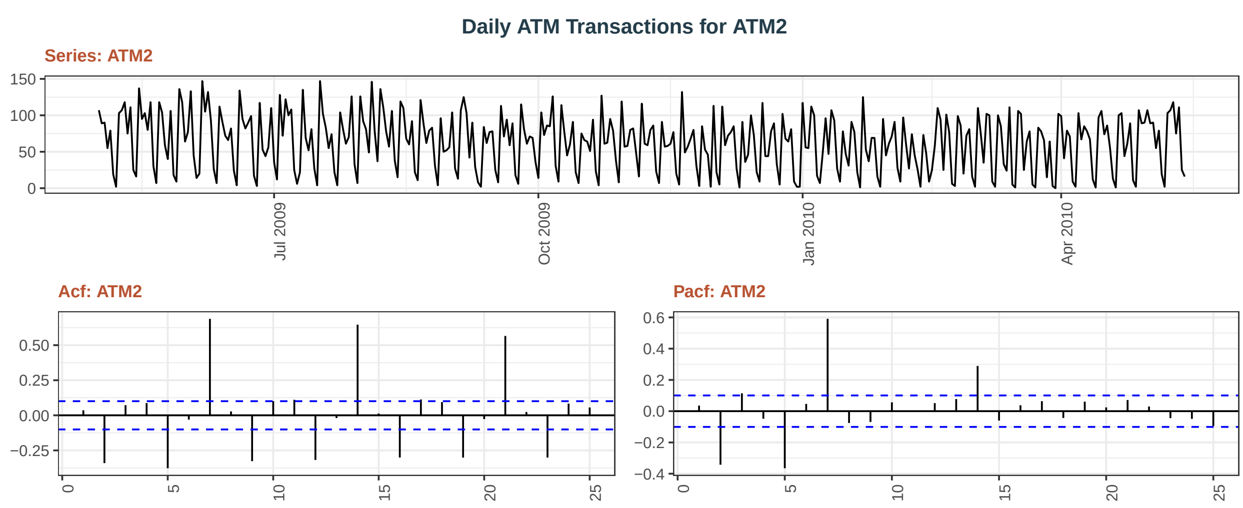
```r
ATM1d %>% diff() %>% ur.kpss() %>% summary()
```

```
FALSE
FALSE #######################
FALSE # KPSS Unit Root Test #
FALSE #######################
FALSE
FALSE Test is of type: mu with 5 lags.
FALSE
FALSE Value of test-statistic is: 0.011
FALSE
FALSE Critical value for a significance level of:
FALSE                 10pct  5pct 2.5pct  1pct
FALSE critical values 0.347 0.463  0.574 0.739
```

Our Acf plot for the ATM1 timeseries shows three large, decreasing lags at 7, 14, and 21. This confirms our assumption about seasonality within our observed data. Our data is non-stationary and should be differentiated in order to forecast the data using a seasonal ARIMA model.

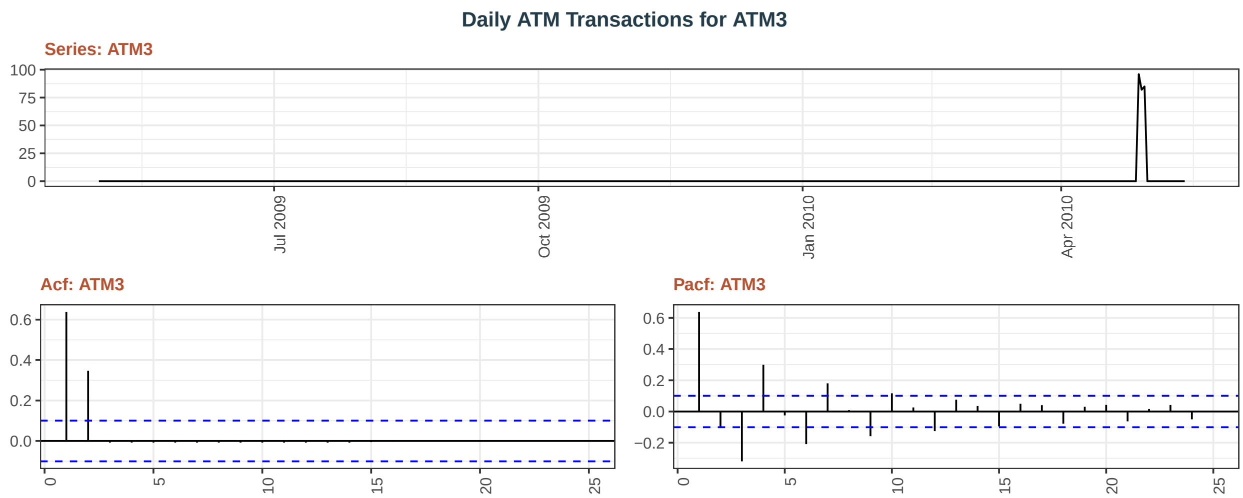
ATM2

```r
#subset data
ATM2 <- atm_ts[,2]

p1<-autoplot(ATM2)+ labs(title="Series: ATM2")+ theme_bw()+theme()
p2<-ggAcf(ATM2)+ labs(title="Acf: ATM2")+ theme_bw()+theme()
p3<-ggPacf(ATM2)+ labs(title="Pacf: ATM2")+ theme_bw()+theme()

grid.arrange(grob=p1, p2, p3,
             layout_matrix = rbind(c(1,1),c(2,3)),
             top=textGrob("Daily ATM Transactions for ATM2",
                          gp = gpar(fontface = "bold", cex = 1, col = "#233c49")))
```



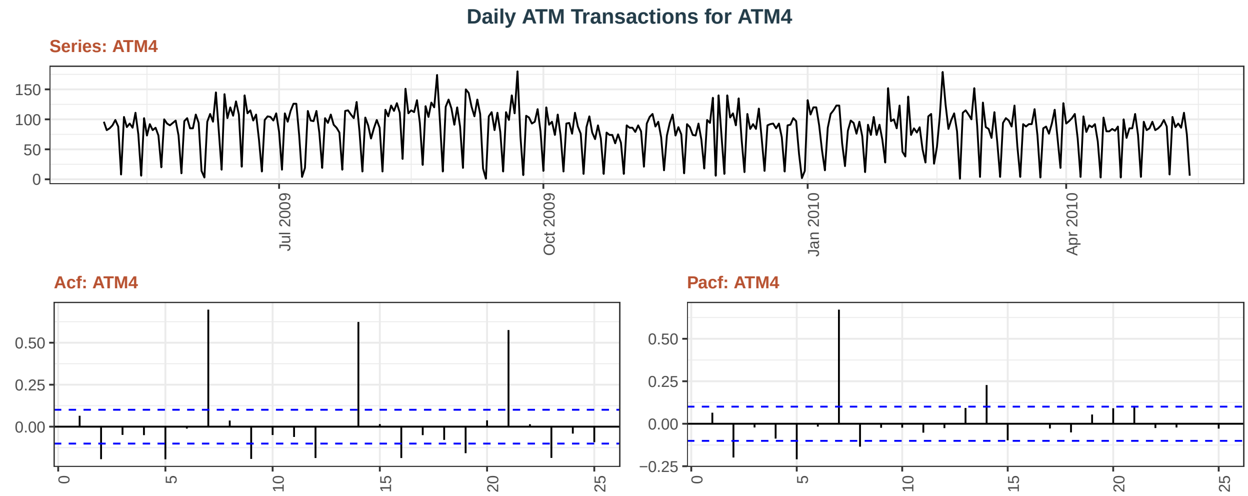
ATM3

```r
#subset data
ATM3 <- atm_ts[,3]

p1<-autoplot(ATM3)+ labs(title="Series: ATM3")+ theme_bw()+theme()
p2<-ggAcf(ATM3)+ labs(title="Acf: ATM3")+ theme_bw()+theme()
p3<-ggPacf(ATM3)+ labs(title="Pacf: ATM3")+ theme_bw()+theme()

grid.arrange(grob=p1, p2, p3,
             layout_matrix = rbind(c(1,1),c(2,3)),
             top=textGrob("Daily ATM Transactions for ATM3",
                          gp = gpar(fontface = "bold", cex = 1, col = "#233c49")))
```



ATM4

```r
#subset data
ATM4 <- atm_ts[,4]

p1<-autoplot(ATM4)+ labs(title="Series: ATM4")+ theme_bw()+theme()
p2<-ggAcf(ATM4)+ labs(title="Acf: ATM4")+ theme_bw()+theme()
p3<-ggPacf(ATM4)+ labs(title="Pacf: ATM4")+ theme_bw()+theme()

grid.arrange(grob=p1, p2, p3,
             layout_matrix = rbind(c(1,1),c(2,3)),
             top=textGrob("Daily ATM Transactions for ATM4",
                          gp = gpar(fontface = "bold", cex = 1, col = "#233c49")))
```



Forecast

Data forecast.

Discussion

Discuss findings.

Part A: Forecasting ATM Withdrawals

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.

Exploration

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

Data exploration.

Forecast

Data forecast.

Discussion

Discuss findings.

Part B: Forecasting Power

Instructions: 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 these to your existing files above - clearly labeled.

Data Exploration

Explore data.

```
power_data <- read_excel("data/ResidentialCustomerForecastLoad-624.xlsx")
```

Data Model

Model data.

Forecast

Data forecast.

Discussion

Discuss findings.

Part C: Forecasting Waterflow

Instructions: 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 test appropriate assumptions and forecast a week forward with confidence bands (80 and 95%). Add these to your existing files above - clearly labeled.

Data Exploration

```r
pipe1_data <- read_excel("data/Waterflow_Pipe1.xlsx")
pipe2_data <- read_excel("data/Waterflow_Pipe2.xlsx")
```

Time-Based Sequence

Create time-based sequence.

Forecast

Data forecast.

Discussion

Discuss findings.