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")
```

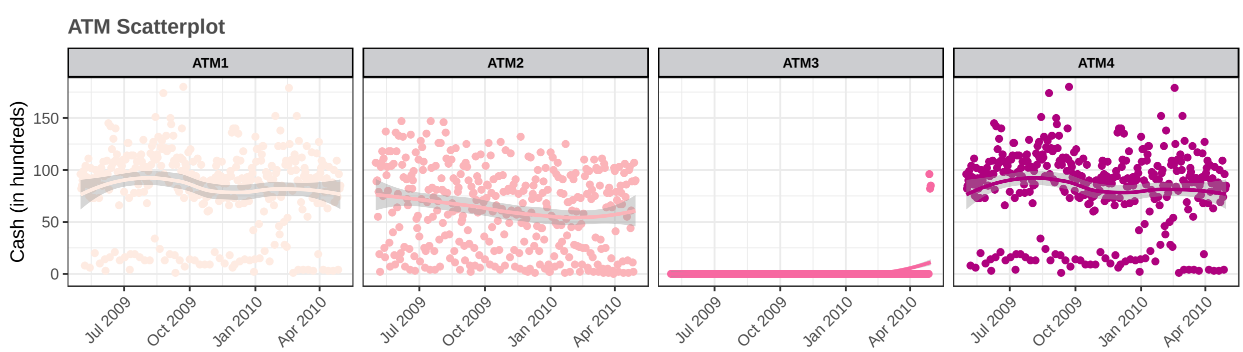
1 Part A

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

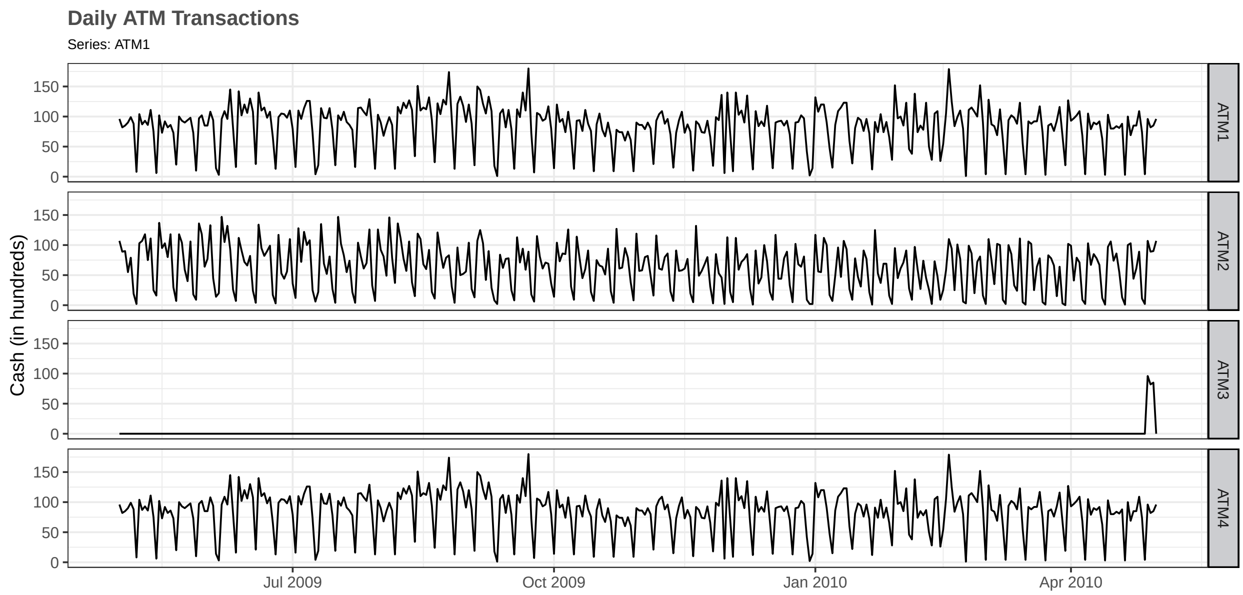
Through data exploration, we identified that the original data file contained `NA` values in our `ATM` and `Cash` columns for 14 observations in May 2010. We removed these missing values and transformed the dataset into a wide format. Our cleaned dataframe was then converted into a timeseries format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



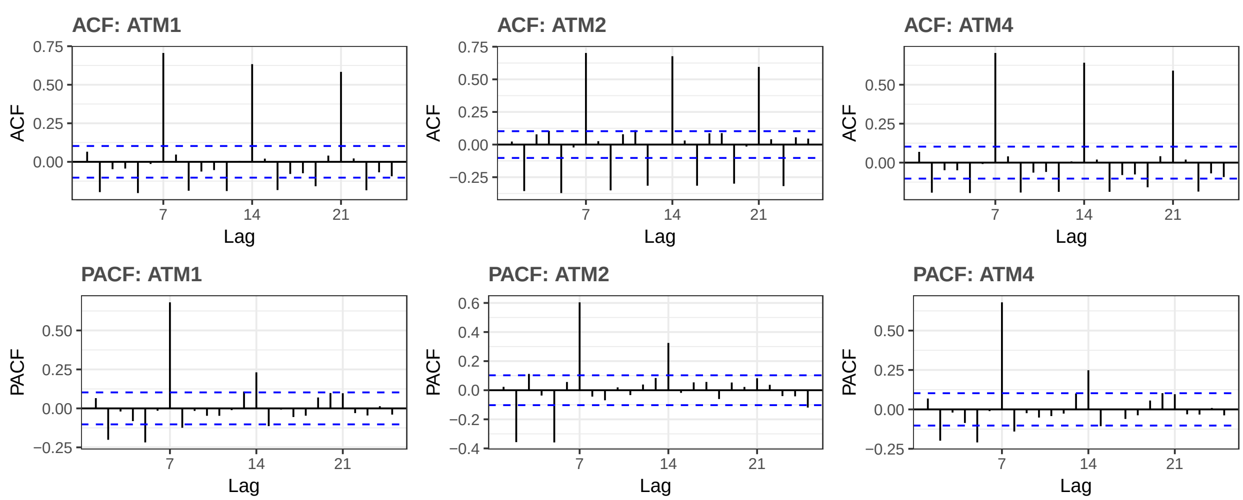
1.2 Timeseries Plots

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 differencing is required on all three series.

Table 1.1: KPSS unit root test

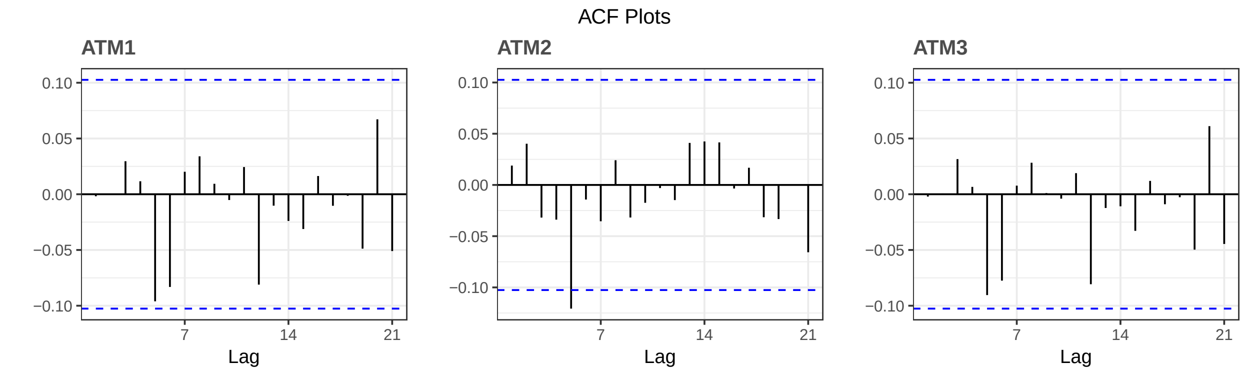
ATM	No-Diff	Diff-1
ATM1	0.4967	0.0219
ATM2	2.0006	0.016
ATM4	0.5182	0.0211

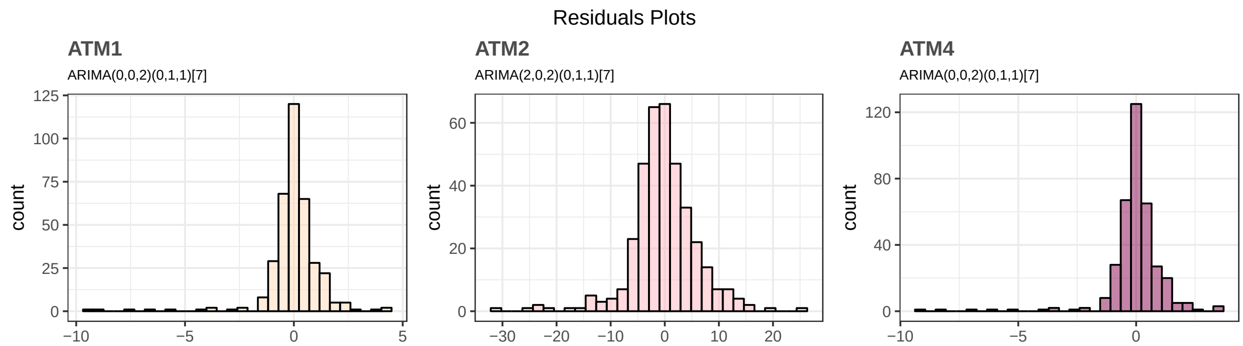
1.3.1 Modeling

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$(0, 0, 2)(0, 1, 1)_7$

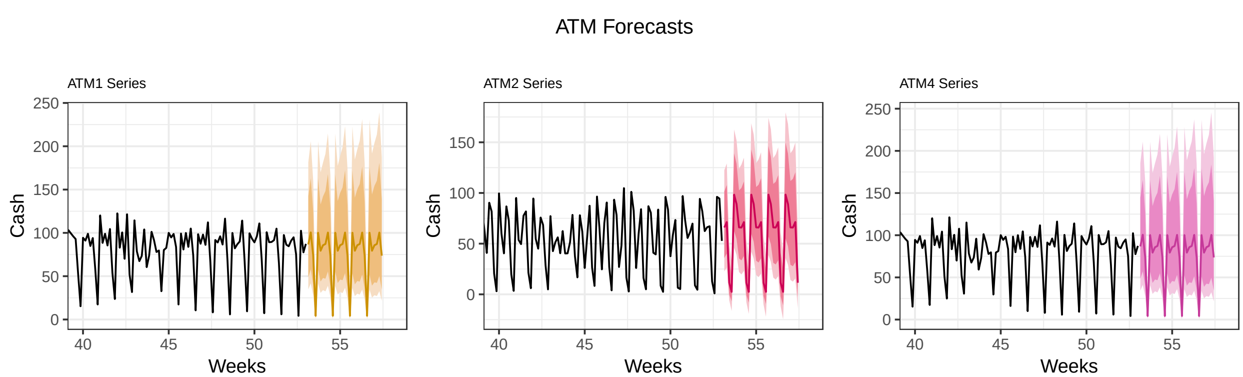
The following ACF plots show us that our differentiated data is now stationary. Further, the residual histograms follow a relatively normal distribution, which confirms that the models adequately fits the observed data.





1.4 Forecast

Finally, we applied a forecast to each series for 31 days, roughly 4.5 weeks, for May 2010. The numeric forecasts can be viewed in a table output 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	48.9373270	142.63088	34.8072635	181.28052
53.28571	100.569238	57.9060340	163.01809	41.7309116	205.83358
53.42857	73.710292	40.2207631	124.43478	27.9573875	159.92833
53.57143	4.229029	1.0444416	11.78734	0.3790053	18.47789
53.71429	100.159253	57.4341570	162.86294	41.2782071	205.92112
53.85714	79.346733	43.8343395	132.71680	30.7254661	169.88926
54.00000	85.739040	47.9707640	142.04415	33.9140346	181.07343
54.14286	87.179762	47.0863041	148.29967	32.5021690	191.23020
54.28571	100.388113	55.5047106	167.80585	38.9277032	214.74786
54.42857	73.710292	38.6103114	128.25525	26.0958862	167.00065
54.57143	4.229029	0.9395260	12.45907	0.3062002	19.92243
54.71429	100.159253	55.3339619	167.52378	38.7869929	214.44281
54.85714	79.346733	42.1172223	136.72351	28.7277449	177.28383
55.00000	85.739040	46.1342721	146.25702	31.7632090	188.82400
55.14286	87.179762	45.3716826	152.41132	30.5286982	198.85654
55.28571	100.388113	53.5677312	172.30784	36.6707415	223.05041
55.42857	73.710292	37.1332203	131.94395	24.4231861	173.89291
55.57143	4.229029	0.8477857	13.11939	0.2474723	21.36255
55.71429	100.159253	53.4026095	172.01735	36.5374994	222.73143
55.85714	79.346733	40.5412620	140.59066	26.9304329	184.48671
56.00000	85.739040	44.4476142	150.32166	29.8257262	196.36992
56.14286	87.179762	43.7858440	156.39578	28.7372821	206.30952
56.28571	100.388113	51.7740714	176.66809	34.6174783	231.15792
56.42857	73.710292	35.7688166	135.52157	22.9086344	180.63578
56.57143	4.229029	0.7669295	13.77059	0.1998318	22.80125
56.71429	100.159253	51.6140616	176.36969	34.4909377	230.82578
56.85714	79.346733	39.0846057	144.34010	25.3010893	191.53034
57.00000	85.739040	42.8876250	154.26122	28.0671555	203.74563
57.14286	87.179762	42.3105860	160.27125	27.1009443	213.61613
57.28571	100.388113	50.1035460	180.90679	32.7379310	239.10038
57.42857	73.710292	34.5011527	139.00405	21.5286686	187.25280

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
53.71429	98.339706	62.6034959	139.16588	46.008826	162.64459
53.85714	89.060722	54.6366990	128.76018	38.844760	151.69804
54.00000	66.068460	35.4504248	102.54649	22.039305	123.94602
54.14286	65.906717	33.5237207	104.96616	19.613265	128.00413
54.28571	71.300878	37.8147203	111.30462	23.208758	134.80084
54.42857	11.465053	-0.6978901	36.62987	-7.713532	53.43919
54.57143	2.455775	-8.1248830	21.53517	-19.241538	35.94422
54.71429	98.360266	59.8144210	142.91141	42.183235	168.68027
54.85714	89.077622	51.9773722	132.39941	35.242609	157.58056
55.00000	66.045852	33.1396694	105.85915	19.076011	129.37285
55.14286	65.902587	31.4119357	108.11314	16.955948	133.16265
55.28571	71.323506	35.6381379	114.54018	20.410600	140.07420
55.42857	11.461937	-1.3413376	38.81996	-9.760929	57.25502
55.57143	2.450060	-9.4745732	23.36433	-21.988596	39.28277
55.71429	98.374495	57.2973343	146.36159	38.780290	174.26039
55.85714	89.089085	49.5803927	135.75469	32.049215	163.02419
56.00000	66.030297	31.0737827	108.93066	16.494819	134.41881
56.14286	65.899881	29.5008716	111.05378	14.617414	137.99752
56.28571	71.339019	33.6587348	117.55610	17.928031	145.00845
56.42857	11.459739	-2.0528650	40.88958	-11.798967	60.87264
56.57143	2.446161	-10.7848065	25.10773	-24.640915	42.47002
56.71429	98.384342	54.9907038	149.58568	35.706429	179.49153
56.85714	89.096857	47.3867226	138.89245	29.175043	168.13139
57.00000	66.019595	29.1975684	111.81596	14.214813	139.17093
57.14286	65.898112	27.7495171	113.83169	12.537521	142.57729
57.28571	71.349653	31.8372509	120.39951	15.701785	149.67600
57.42857	11.458191	-2.8069332	42.86353	-13.824563	64.33275

Table 1.4: ATM4 Forecast

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
53.14286	86.714879	48.4833222	144.69401	34.4084146	185.45461
53.28571	100.581689	57.2295950	165.56192	41.0835082	210.92017
53.42857	73.645362	39.8822791	125.94463	27.7136726	163.18586
53.57143	4.221433	1.1652846	11.35003	0.4840215	17.69213
53.71429	100.159422	56.7554394	165.39669	40.6339225	211.01094
53.85714	79.341721	43.4564455	134.51304	30.4249539	173.62248
54.00000	85.778198	47.5307114	144.12943	33.5336578	185.30067
54.14286	87.218338	46.6724830	150.52084	32.1753416	195.81061
54.28571	100.395110	54.8758085	170.47265	38.3681897	220.18913
54.42857	73.645362	38.3197899	129.82279	25.9245335	170.45814
54.57143	4.221433	1.0633596	11.96752	0.4069650	19.02993
54.71429	100.159422	54.7048087	170.17312	38.2291112	219.85960
54.85714	79.341721	41.7871128	138.59132	28.5019577	181.24714
55.00000	85.778198	45.7420241	148.42877	31.4604964	193.31376
55.14286	87.218338	45.0004935	154.72823	30.2708579	203.72069
55.28571	100.395110	52.9828581	175.09732	36.1876286	228.83531
55.42857	73.645362	36.8851955	133.57555	24.3147396	177.56475
55.57143	4.221433	0.9733641	12.57513	0.3430812	20.36576
55.71429	100.159422	52.8175347	174.78874	36.0559480	228.49057
55.85714	79.341721	40.2535029	142.53639	26.7698277	188.69467
56.00000	85.778198	44.0977575	152.58623	29.5910200	201.13698
56.14286	87.218338	43.4530658	158.81358	28.5406851	211.47028
56.28571	100.395110	51.2290077	179.58544	34.2027660	237.29975
56.42857	73.645362	35.5590027	137.22241	22.8555483	184.53442
56.57143	4.221433	0.8932856	13.17487	0.2897334	21.70226
56.71429	100.159422	51.0688494	179.26833	34.0777287	236.94056
56.85714	79.341721	38.8349570	146.36891	25.1980635	195.99558
57.00000	85.778198	42.5759587	156.62374	27.8927754	208.80278
57.14286	87.218338	42.0128505	162.79425	26.9592616	219.08495
57.28571	100.395110	49.5949518	183.95633	32.3850894	245.61097
57.42857	73.645362	34.3260932	140.77849	21.5247583	191.38935

R Script

```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
```

```r
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))

# 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=31)
ATM2_fc <- forecast(ATM2_arima,h=31)
ATM4_fc <- forecast(ATM4_arima,h=31)

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

```r
write.csv(ATM4_fc, file="forecasts/ATM4_Forecast.csv")
```