Data624 - Project1

Amit Kapoor

3/28/2021

Contents

Overview

This project includes 3 time series dataset and requires to select best forecasting model for all 3 datasets.

- Part A - ATM Forecast
- Part B - Forecasting Power
- Part C - Waterflow Pipe

Part A - ATM Forecast

The dataset contains cash withdrawals from 4 different ATM machines from May 2009 to Apr 2010. The variable 'Cash' is provided in hundreds of dollars and data is in a single file. Before starting our analysis we will first download the excel from github and then read it through read_excel.

Exploratory Analysis

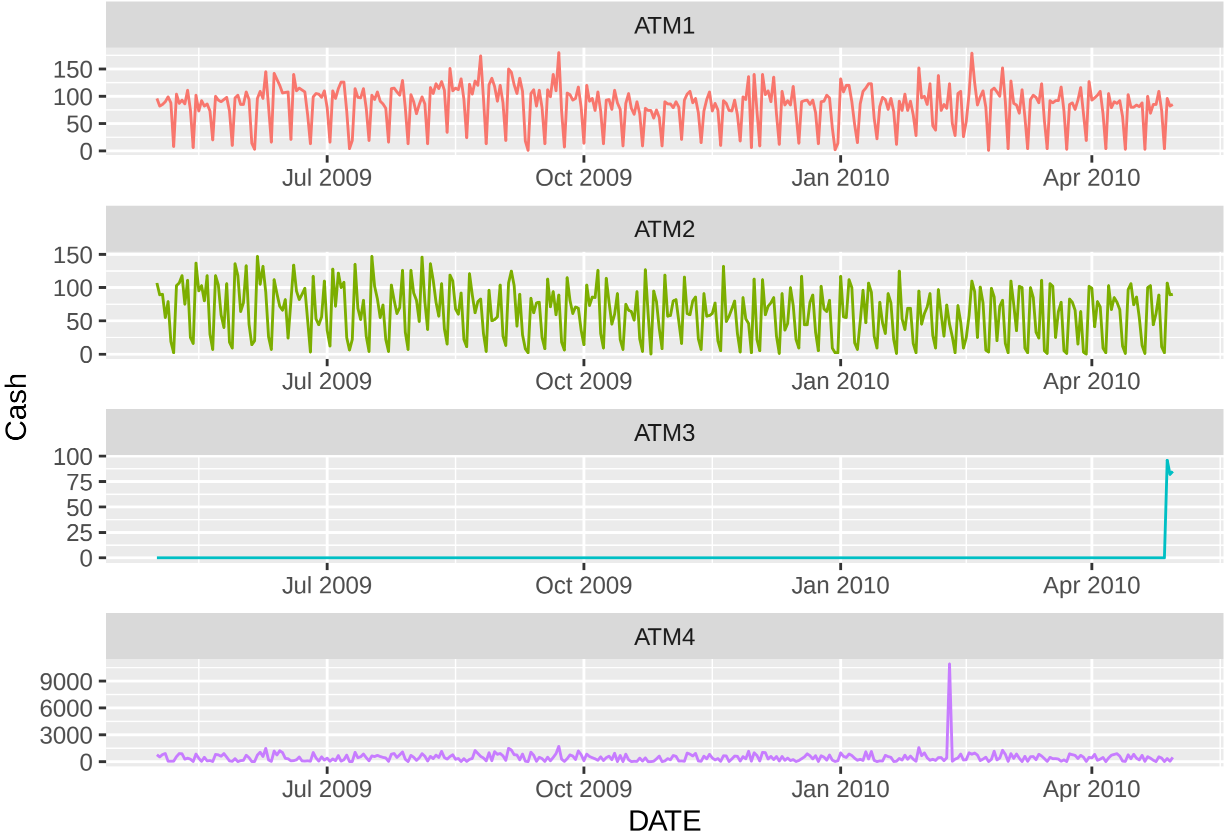
```
temp.file <- tempfile(fileext = ".xlsx")
download.file(url="https://github.com/amit-kapoor/data624/blob/main/Project1/ATM624Data.xlsx?raw=true",
              destfile = temp.file,
              mode = "wb",
              quiet = TRUE)
atm.data <- read_excel(temp.file, skip=0, col_types = c("date","text","numeric"))

glimpse(atm.data)
```

```
## Rows: 1,474
## Columns: 3
## $ DATE <dttm> 2009-05-01, 2009-05-01, 2009-05-02, 2009-05-02, 2009-05-03, 2009~
## $ ATM  <chr> "ATM1", "ATM2", "ATM1", "ATM2", "ATM1", "ATM2", "ATM1", "ATM2", "~
## $ Cash <dbl> 96, 107, 82, 89, 85, 90, 90, 55, 99, 79, 88, 19, 8, 2, 104, 103, ~
```

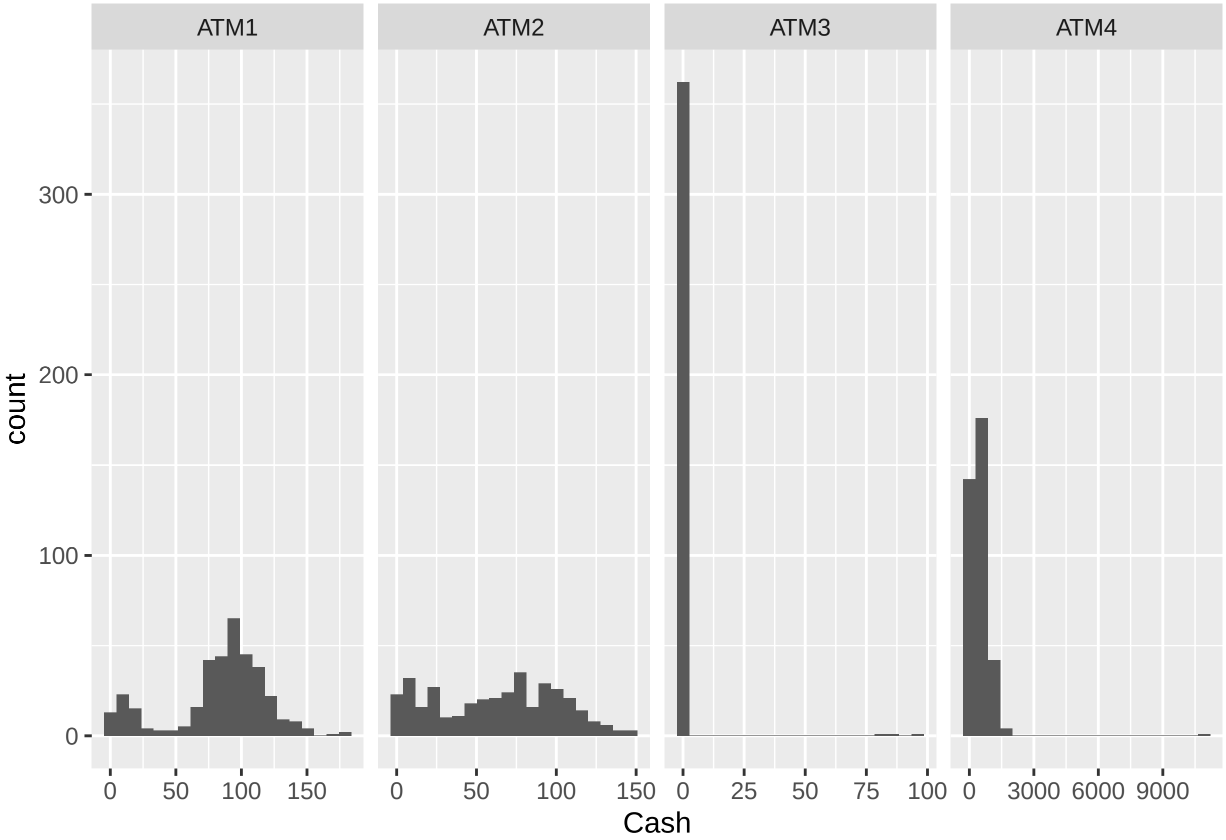
```r
# rows missing values
atm.data[!complete.cases(atm.data),]
```

```
## # A tibble: 19 x 3
##    DATE                ATM    Cash
##    <dttm>              <chr> <dbl>
##  1 2009-06-13 00:00:00 ATM1     NA
##  2 2009-06-16 00:00:00 ATM1     NA
##  3 2009-06-18 00:00:00 ATM2     NA
##  4 2009-06-22 00:00:00 ATM1     NA
##  5 2009-06-24 00:00:00 ATM2     NA
##  6 2010-05-01 00:00:00 <NA>     NA
##  7 2010-05-02 00:00:00 <NA>     NA
##  8 2010-05-03 00:00:00 <NA>     NA
##  9 2010-05-04 00:00:00 <NA>     NA
## 10 2010-05-05 00:00:00 <NA>     NA
## 11 2010-05-06 00:00:00 <NA>     NA
## 12 2010-05-07 00:00:00 <NA>     NA
## 13 2010-05-08 00:00:00 <NA>     NA
## 14 2010-05-09 00:00:00 <NA>     NA
## 15 2010-05-10 00:00:00 <NA>     NA
## 16 2010-05-11 00:00:00 <NA>     NA
## 17 2010-05-12 00:00:00 <NA>     NA
## 18 2010-05-13 00:00:00 <NA>     NA
## 19 2010-05-14 00:00:00 <NA>     NA
```

```r
ggplot(atm.data[complete.cases(atm.data),] , aes(x=DATE, y=Cash, col=ATM )) +
  geom_line(show.legend = FALSE) +
  facet_wrap(~ATM, ncol=1, scales = "free")
```



```
ggplot(atm.data[complete.cases(atm.data),] , aes(x=Cash )) +
  geom_histogram(bins=20) +
  facet_grid(cols=vars(ATM), scales = "free")
```



```
# consider complete cases
atm.comp <- atm.data[complete.cases(atm.data),]
# pivot wider with cols from 4 ATMs and their values as Cash
atm.comp <- atm.comp %>% pivot_wider(names_from = ATM, values_from = Cash)
head(atm.comp)
```

```
## # A tibble: 6 x 5
##   DATE                 ATM1  ATM2  ATM3  ATM4
##   <dttm>              <dbl> <dbl> <dbl> <dbl>
## 1 2009-05-01 00:00:00    96   107     0 777.
## 2 2009-05-02 00:00:00    82    89     0 524.
## 3 2009-05-03 00:00:00    85    90     0 793.
## 4 2009-05-04 00:00:00    90    55     0 908.
## 5 2009-05-05 00:00:00    99    79     0  52.8
## 6 2009-05-06 00:00:00    88    19     0  52.2
```

```
# summary
atm.comp %>% select(-DATE) %>% summary()
```

```
##       ATM1             ATM2             ATM3              ATM4
##  Min.   :  1.00   Min.   :  0.00   Min.   : 0.0000   Min.   :    1.563
##  1st Qu.: 73.00   1st Qu.: 25.50   1st Qu.: 0.0000   1st Qu.:  124.334
##  Median : 91.00   Median : 67.00   Median : 0.0000   Median :  403.839
##  Mean   : 83.89   Mean   : 62.58   Mean   : 0.7206   Mean   :  474.043
##  3rd Qu.:108.00   3rd Qu.: 93.00   3rd Qu.: 0.0000   3rd Qu.:  704.507
##  Max.   :180.00   Max.   :147.00   Max.   :96.0000   Max.   :10919.762
##  NA's   :3        NA's   :2
```

Per above exploratory analysis, all ATMs show different patterns. We would perform forecasting for each

ATM separately.

- ATM1 and ATM2 shows similar pattern (approx.) throughout the time. ATM1 and ATM2 have 3 and 2 missing entries respectively.
- ATM3 appears to become online in last 3 days only and rest of days appears inactive. So tha data available for this ATM is very limited.
- ATM4 requires replacement for outlier and we can assume that one day spike of cash withdrawal is unique. It has an outlier showing withdrawl amount 10920.

Data Cleaning

```
atm.ts <- ts(atm.comp %>% select(-DATE))
head(atm.ts)
```

```
## Time Series:
## Start = 1
## End = 6
## Frequency = 1
##   ATM1 ATM2 ATM3      ATM4
## 1   96  107    0 776.99342
## 2   82   89    0 524.41796
## 3   85   90    0 792.81136
## 4   90   55    0 908.23846
## 5   99   79    0  52.83210
## 6   88   19    0  52.20845
```

```
atm.ts.cln <- sapply(X=atm.ts, tsclean)
atm.ts.cln %>% summary()
```

```
##       ATM1             ATM2             ATM3              ATM4         
##  Min.   :  1.00   Min.   :  0.00   Min.   : 0.0000   Min.   :   1.563  
##  1st Qu.: 73.00   1st Qu.: 26.00   1st Qu.: 0.0000   1st Qu.: 124.334  
##  Median : 91.00   Median : 67.00   Median : 0.0000   Median : 402.770  
##  Mean   : 84.15   Mean   : 62.59   Mean   : 0.7206   Mean   : 444.757  
##  3rd Qu.:108.00   3rd Qu.: 93.00   3rd Qu.: 0.0000   3rd Qu.: 704.192  
##  Max.   :180.00   Max.   :147.00   Max.   :96.0000   Max.   :1712.075
```

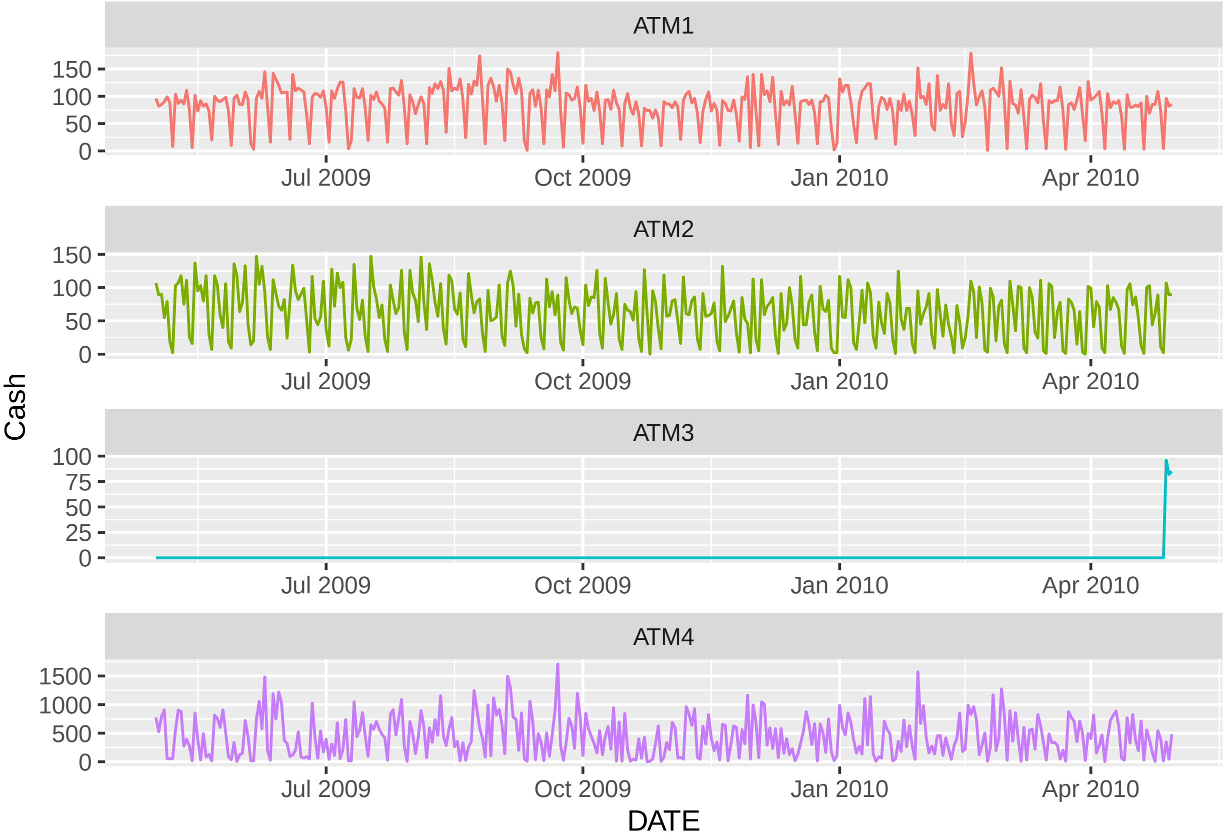
```
# convert into data frame, pivot longer , arrange by ATM and bind with dates
atm.new <- as.data.frame(atm.ts.cln) %>%
  pivot_longer(everything(), names_to = "ATM", values_to = "Cash") %>%
  arrange(ATM)

atm.new <- cbind(DATE = seq(as.Date("2009-05-1"), as.Date("2010-04-30"), length.out=365),
                 atm.new)
head(atm.new)
```

```
##         DATE  ATM Cash
## 1 2009-05-01 ATM1   96
## 2 2009-05-02 ATM1   82
## 3 2009-05-03 ATM1   85
## 4 2009-05-04 ATM1   90
## 5 2009-05-05 ATM1   99
## 6 2009-05-06 ATM1   88
```

```
#library(xlsx)
```

```
#write.xlsx(atm.new, 'atmnew.xlsx', sheetName = "Sheet1", col.names = TRUE, row.names = TRUE, append = 
```

```
ggplot(atm.new , aes(x=DATE, y=Cash, col=ATM )) +
  geom_line(show.legend = FALSE) +
  facet_wrap(~ATM, ncol=1, scales = "free")
```



Part B - Forecasting Power

```
download.file(
  url="https://github.com/amit-kapoor/data624/blob/main/Project1/ResidentialCustomerForecastLoad-624.xl
  destfile = temp.file,
  mode = "wb",
  quiet = TRUE)
power.data <- read_excel(temp.file, skip=0, col_types = c("numeric","text","numeric"))

head(power.data)
```

```
## # A tibble: 6 x 3
##   CaseSequence `YYYY-MMM`     KWH
##          <dbl> <chr>        <dbl>
## 1          733 1998-Jan   6862583
## 2          734 1998-Feb   5838198
## 3          735 1998-Mar   5420658
## 4          736 1998-Apr   5010364
## 5          737 1998-May   4665377
## 6          738 1998-Jun   6467147
```

Part C - Waterflow Pipe

```
download.file(url="https://github.com/amit-kapoor/data624/blob/main/Project1/Waterflow_Pipe1.xlsx?raw=t
              destfile = temp.file,
              mode = "wb",
              quiet = TRUE)
pipe1.data <- read_excel(temp.file, skip=0, col_types = c("date","numeric"))

download.file(url="https://github.com/amit-kapoor/data624/blob/main/Project1/Waterflow_Pipe2.xlsx?raw=t
              destfile = temp.file,
              mode = "wb",
              quiet = TRUE)

pipe2.data <- read_excel(temp.file, skip=0, col_types = c("date","numeric"))
head(pipe1.data)
```

```
## # A tibble: 6 x 2
##   `Date Time`         WaterFlow
##   <dttm>                  <dbl>
## 1 2015-10-23 00:24:06     23.4
## 2 2015-10-23 00:40:02     28.0
## 3 2015-10-23 00:53:51     23.1
## 4 2015-10-23 00:55:40     30.0
## 5 2015-10-23 01:19:17      6.00
## 6 2015-10-23 01:23:58     15.9
```