DATA624 - Project1

Glen Dale Davis

2023-10-14

Packages:

```
library(fpp3)
library(RColorBrewer)
library(knitr)
library(pracma)
library(cowplot)
library(readxl)
library(httr)
```

Part A:

Data Preparation:

We load transaction data for four ATMS from May 2009 to April 2010.

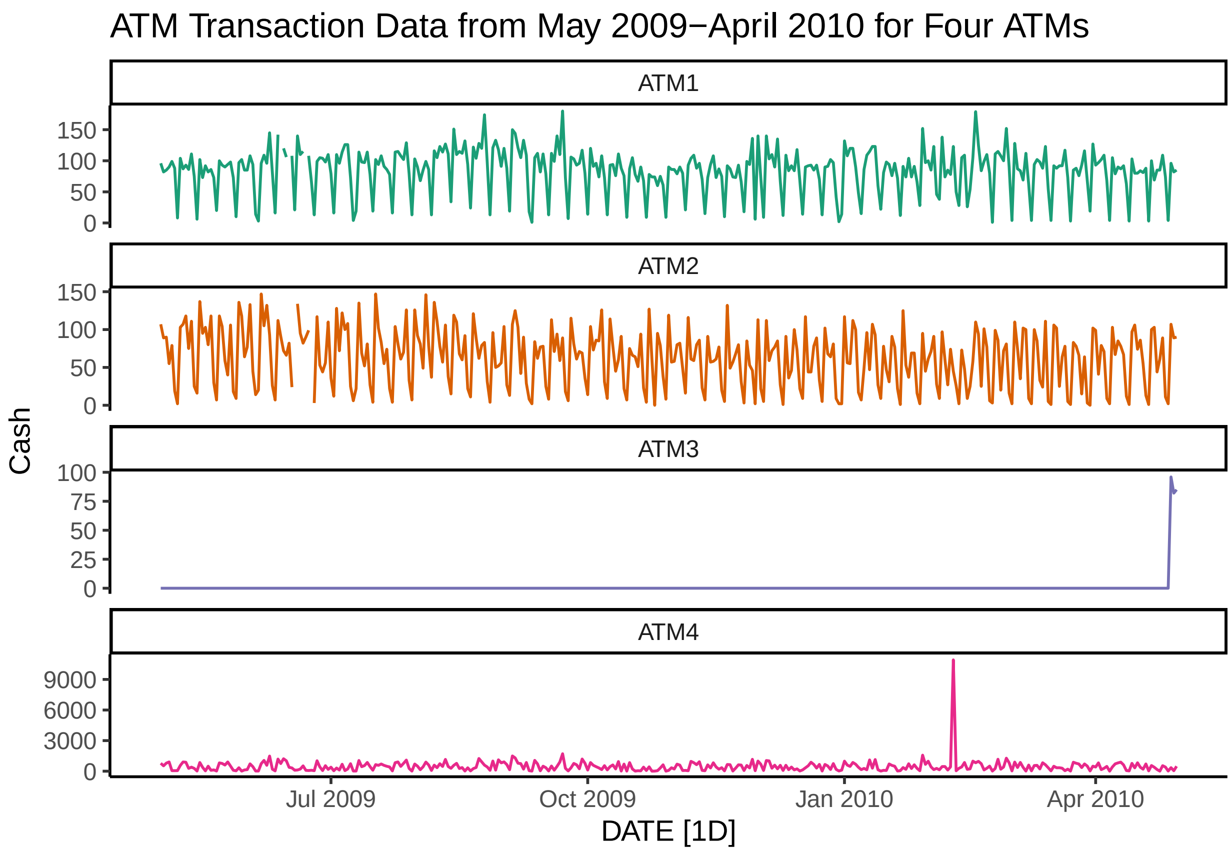
```
my_url <- "https://github.com/geedoubledee/data624_project1/raw/main/ATM624Data.xlsx"
col_types <- c("date", "text", "numeric")
temp <- tempfile(fileext = ".xlsx")
req <- GET(my_url, authenticate(Sys.getenv("GITHUB_PAT"), ""),
           write_disk(path = temp))
atm <- readxl::read_excel(temp, col_types = col_types)
```

We coerce the `DATE` variable to class *Date* and the `ATM` variable to class *Factor*. We remove observations that contain neither `ATM` nor `CASH` data. (These empty observations all occur during the period we will be forecasting for: May 2010.)

```
levels = c("ATM1", "ATM2", "ATM3", "ATM4")
atm <- atm |>
    mutate(DATE = as.Date(DATE),
           ATM = factor(ATM, levels = levels, ordered = TRUE)) |>
    filter(!is.na(ATM))
```

We create and plot the time series with subplots for each ATM, and we print a summary of 0 or `NA` values for each ATM.

```r
theme_set(theme_classic())
palette <- brewer.pal(n = 8, name = "Dark2")
atm_colors <- palette[1:4]
names(atm_colors) <- levels
atm_ts <- atm |>
    as_tsibble(index = DATE, key = ATM)
p1 <- atm_ts |>
    autoplot(Cash) +
    facet_wrap(~ ATM, scales = "free_y" , ncol = 1) +
    scale_color_brewer(palette = "Dark2") +
    theme(legend.position = "none") +
    labs(title = "ATM Transaction Data from May 2009-April 2010 for Four ATMs")
p1
```



```r
na_summary <- atm_ts |>
    as_tibble() |>
    filter(is.na(Cash) | Cash == 0) |>
    group_by(ATM, Cash) |>
    summarize(Count = n())
```

```
## ‘summarise()‘ has grouped output by ’ATM’. You can override using the ‘.groups‘
## argument.
```

```
knitr::kable(na_summary)
```

ATM	Cash	Count
ATM1	NA	3
ATM2	0	2
ATM2	NA	2
ATM3	0	362

We can immediately see missing value and outlier issues with the data that we will need to address. First, we will handle ATM3's data issues, as those need to be addressed differently from the other ATMs' data issues.

The distribution of `Cash` is degenerate for ATM3, as 99% of values are 0. The most likely interpretation of these values is that the machine was not operating during the period prior to the first non-zero recorded value. Since we will be forecasting for each ATM separately, and we therefore cannot exclude ATM3 from our analysis, we will treat these 0 values as missing and use Next Observation Carried Backward (NOCB) to fill them. (This technique is the same as Last Observation Carried Forward (LOCF), except it is performed in the opposite direction.)

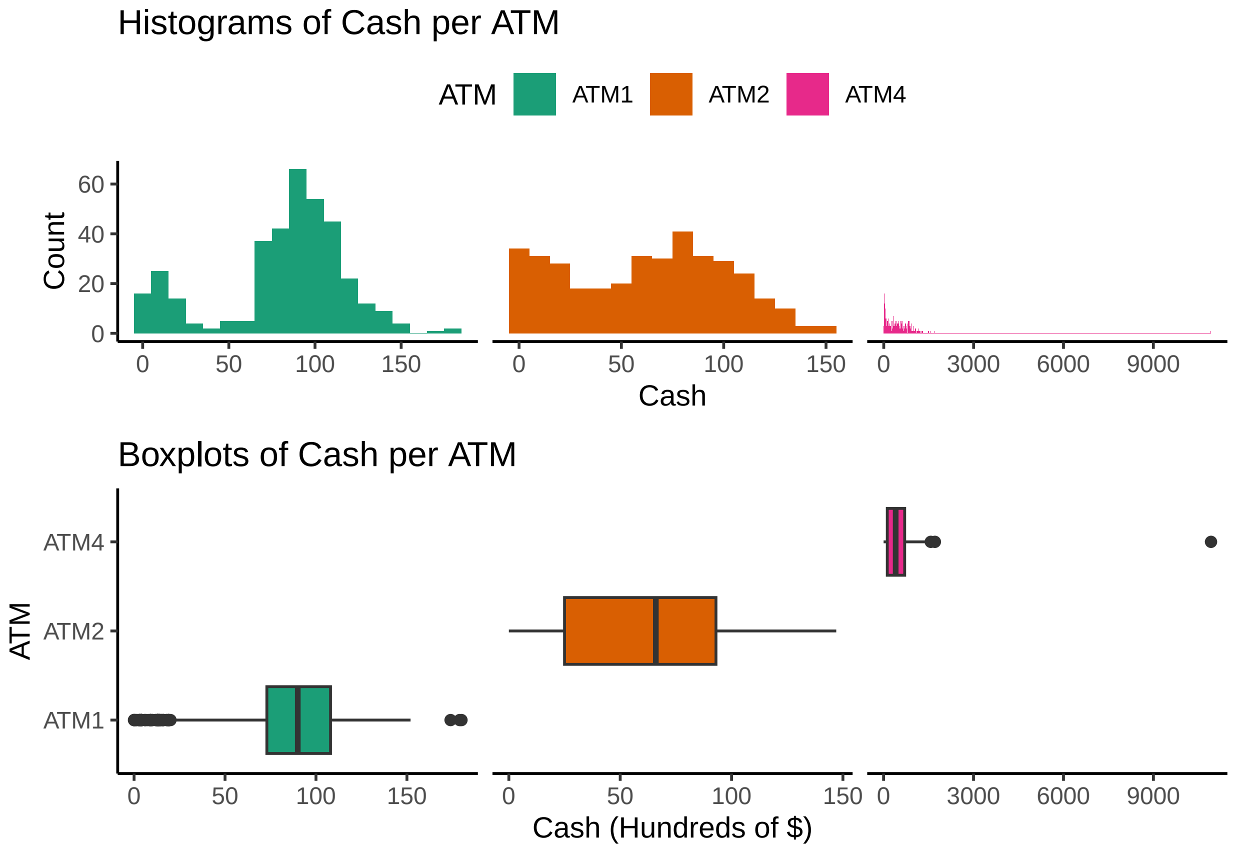
```
atm3 <- atm_ts |>
    filter(ATM == "ATM3")
atm3$Cash[atm3$Cash == 0] <- NA
atm3 <- atm3 |>
    fill(Cash, .direction = "up")
atm_ts <- atm_ts|>
    filter(ATM != "ATM3") |>
    bind_rows(atm3)
```

Next we take a closer look at ATMs 1, 2, and 4.

Since ATMs 1 and 2 have both 0 and `NA` values, the `NA` values will first be converted to 0 values. It is reasonable that these observations correctly indicate that no transactions occurred at these ATMs on those days.

```
atm_ts_na <- atm_ts |>
    filter(is.na(Cash)) |>
    mutate(Cash = 0)
atm_ts <- atm_ts|>
    filter(!is.na(Cash)) |>
    bind_rows(atm_ts_na)
p2a <- atm_ts |>
    filter(ATM != "ATM3") |>
    ggplot(aes(x = Cash, fill = ATM)) +
    geom_histogram(binwidth = 10) +
    facet_wrap(~ ATM, scales = "free_x") +
    scale_fill_manual(values = atm_colors) +
    theme(legend.position = "top",
          strip.background = element_blank(),
          strip.text.x = element_blank()) +
    labs(title = "Histograms of Cash per ATM",
         y = "Count")
p2b <- atm_ts |>
```

```
    filter(ATM != "ATM3") |>
    ggplot(aes(x=Cash, y=ATM, fill=ATM)) +
    geom_boxplot() +
    facet_grid(~ ATM, scales = "free_x") +
    labs(title="Boxplots of Cash per ATM",x="Cash (Hundreds of $)", y = "ATM") +
    scale_fill_manual(values = atm_colors) +
    theme(strip.background = element_blank(),
          strip.text.x = element_blank(),
          legend.position = "none")
p2 <- plot_grid(p2a, p2b, ncol = 1, align = "v", axis = "l")
p2
```



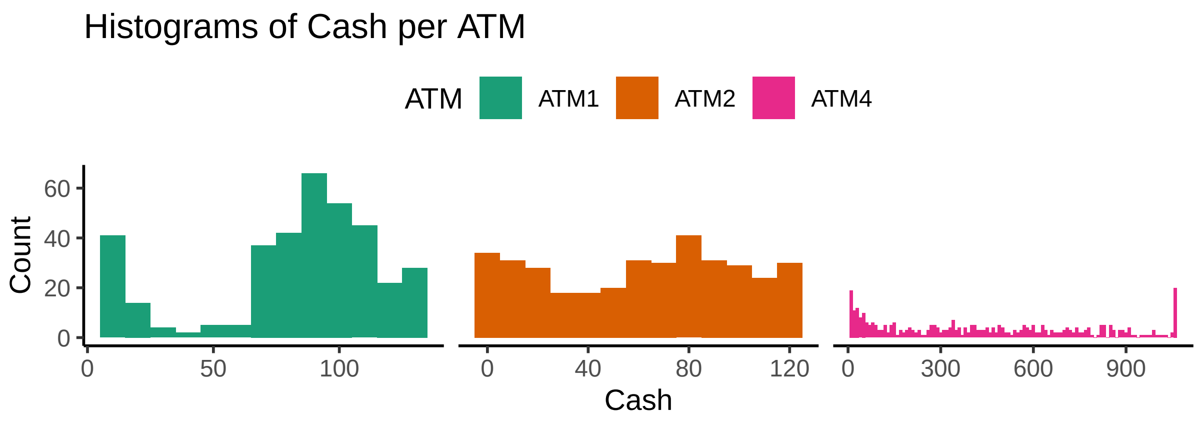
One `Cash` value is over 6 times larger than any other value for ATM4 or any other ATM in this dataset. Our understanding is that ATMs operated in retail locations (i.e. gas stations or hotels) typically hold a max of $\$20,000$, and ATMs operated in banks can hold up to $\$200,000$. We will replace this extreme value and other outliers for ATMs 1, 2, and 4 in the dataset by winsorizing them. By winsorizing, we replace any value of a variable above or below percentile k with the value of the k^th percentile itself. A k of 5 is standard, so we will replace outliers below the 5th percentile with the value of the 5th percentile, and we will replace outliers above the 95th percentile with the value of the 95th percentile.

However, our winsorization method will apply to all values below the 5th percentile, including any 0 values.

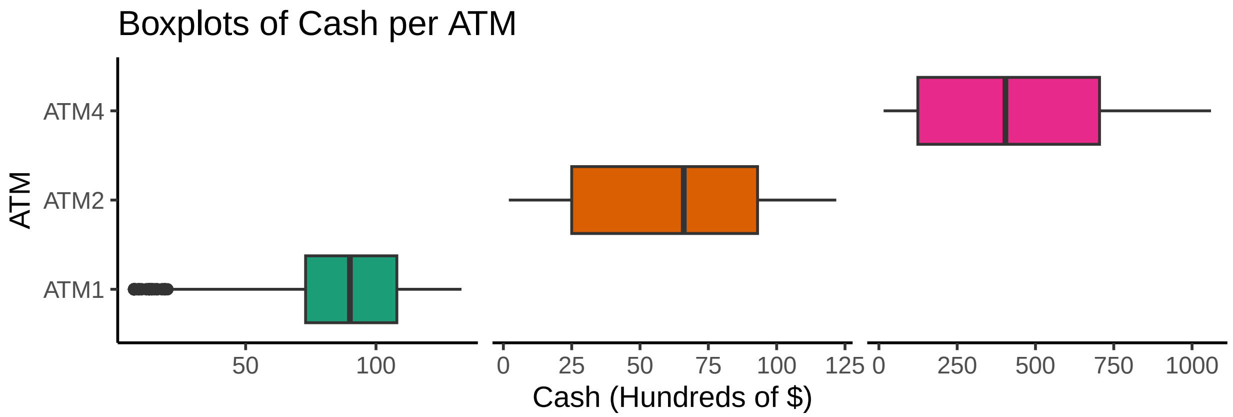
```
atm_ts <- atm_ts |>
    group_by(ATM) |>
    mutate(Cash = case_when(
        Cash > quantile(Cash, probs = 0.95) ~ quantile(Cash, probs = 0.95),
        Cash < quantile(Cash, probs = 0.05) ~ quantile(Cash, probs = 0.05),
        TRUE ~ Cash))
```

Now we can confirm whether the outliers in the distributions for ATMS 1, 2, and 4 have been removed.

```
p3a <- atm_ts |>
    filter(ATM != "ATM3") |>
    ggplot(aes(x = Cash, fill = ATM)) +
    geom_histogram(binwidth = 10) +
    facet_wrap(~ ATM, scales = "free_x") +
    scale_fill_manual(values = atm_colors) +
    theme(legend.position = "top",
          strip.background = element_blank(),
          strip.text.x = element_blank()) +
    labs(title = "Histograms of Cash per ATM",
         y = "Count")
p3b <- atm_ts |>
    filter(ATM != "ATM3") |>
    ggplot(aes(x=Cash, y=ATM, fill=ATM)) +
    geom_boxplot() +
    facet_grid(~ ATM, scales = "free_x") +
    labs(title="Boxplots of Cash per ATM",
         x="Cash (Hundreds of $)",
         y = "ATM") +
    scale_fill_manual(values = atm_colors) +
    theme(strip.background = element_blank(),
          strip.text.x = element_blank(),
          legend.position = "none")
p3 <- plot_grid(p3a, p3b, ncol = 1, align = "v", axis = "l")
p3
```



Histograms of Cash per ATM



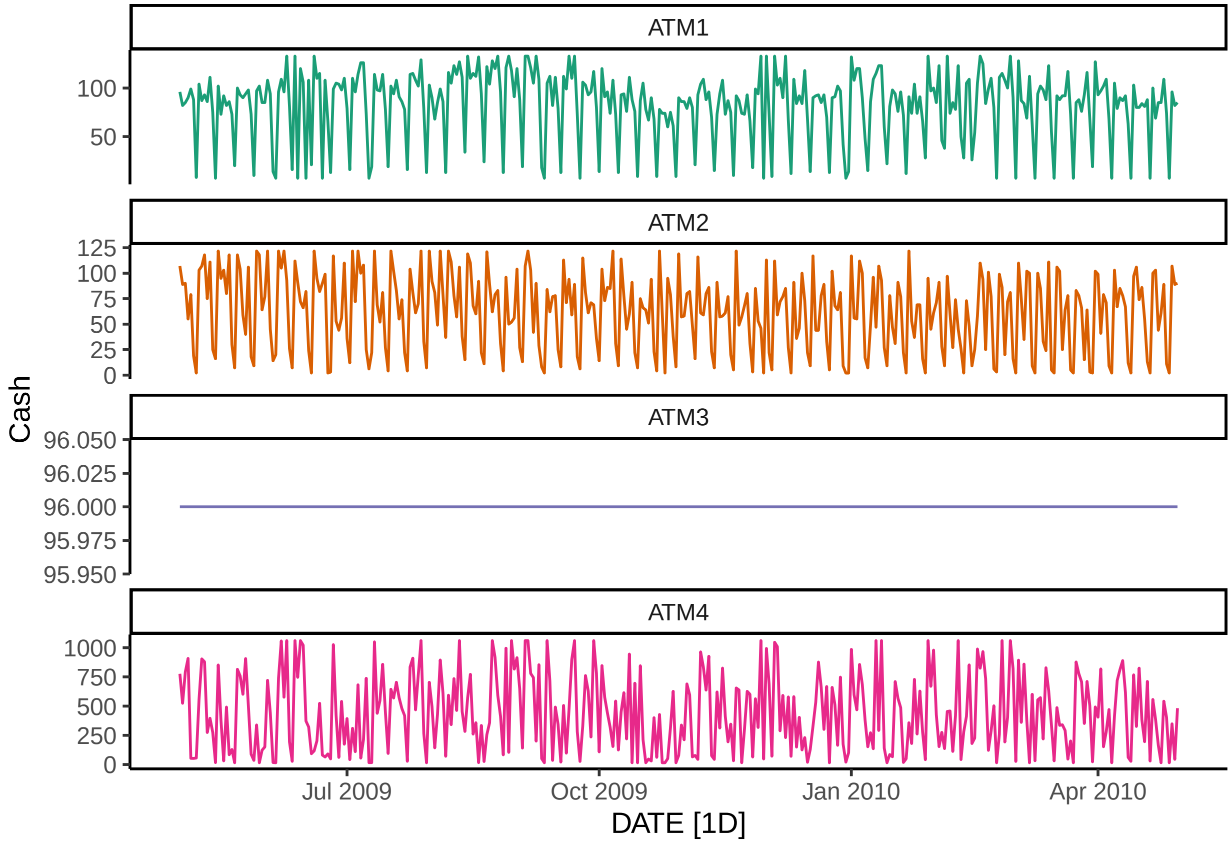
Boxplots of Cash per ATM

All the outliers for ATMs 2 and 4 have been replaced. ATM1 still contains outliers on the low end.

```
p5 <- atm_ts |>
    autoplot(Cash) +
    facet_wrap(~ ATM, scales = "free_y" , ncol = 1) +
    scale_color_brewer(palette = "Dark2") +
    theme(legend.position = "none")
```

```
## ‘mutate_if()‘ ignored the following grouping variables:
## * Column ‘ATM‘
```

```
p5
```



```
#p6 <- atm_ts |>
    #gg_season(Cash, period = "week") +
    #facet_wrap(~ ATM, scales = "free_y" , ncol = 1) +
    #scale_color_brewer(palette = "Dark2") +
    #theme(strip.background = element_blank(),
          #strip.text.x = element_blank())
#p6
```

Part B:

Part C: