Project 1

Paul Perez

4/10/2021

Part A - ATM Forecast

Our goal is for Part A of project 1 to forecast how much cash is taken out of the 4 different ATM machines for May 2010. Given the excel file containing all of our data, there are three columns; `DATE`, `ATM`, and `Cash`. We have to explore the dat and determine the best way to forecast, with little direction.

Data Collection

As collect the data, we'll explore the format of the data, types of variables, and

```
atm <- read_excel("ATM624Data.xlsx")
df.atm <- data.frame(atm)
```

Data Cleanse

```
head(df.atm)
```

```
##    DATE  ATM Cash
## 1 39934 ATM1   96
## 2 39934 ATM2  107
## 3 39935 ATM1   82
## 4 39935 ATM2   89
## 5 39936 ATM1   85
## 6 39936 ATM2   90
```

```
dim(df.atm)
```

```
## [1] 1474    3
```

Instantly, we could see that the date format needs to be converted to an actual date format, and the shape of the dataset is 1474, 3 meaning there are 1474 records across 3 columns. We can convert that using the base R cast of `as.Date()` function. This information was source from the below stackoverflow link. How to convert Excel date format to proper date in R

```
df.atm$DATE <- as.Date(df.atm$DATE, origin = "1899-12-30")
head(df.atm)
```

```
##         DATE  ATM Cash
## 1 2009-05-01 ATM1   96
## 2 2009-05-01 ATM2  107
## 3 2009-05-02 ATM1   82
## 4 2009-05-02 ATM2   89
## 5 2009-05-03 ATM1   85
## 6 2009-05-03 ATM2   90
```

```
dim(df.atm)
```

```
## [1] 1474    3
```

Now that we have a dataset with actual dates, we can further evaluate the whole dataset.

```
summary(df.atm)
```

```
##       DATE                 ATM                Cash        
##  Min.   :2009-05-01   Length:1474        Min.   :    0.0  
##  1st Qu.:2009-08-01   Class :character   1st Qu.:    0.5  
##  Median :2009-11-01   Mode  :character   Median :   73.0  
##  Mean   :2009-10-31                      Mean   :  155.6  
##  3rd Qu.:2010-02-01                      3rd Qu.:  114.0  
##  Max.   :2010-05-14                      Max.   :10919.8  
##                                          NA's   :19
```

We can see that the `DATE` column ranges from May 1st, 2010 through May 14th, 2010. The `ATM` column has a length of 1474 confirming the `dim()` function earlier. Additionally, looking at the dataframe preview above using the `head()` function, we can see that there are multiple ATM's in the column. Our third column, `Cash`, has a minimum value of 0.0 and a maximum value of 10919.8. There are 19 NULL values that we can handle when we preprocess the data.

Complete Dataframe Analysis

First, we can evaluate the full dataframe, inclusive of all 4 ATM's data. Afterwards, we can evaluate the subset of the dataframe for each ATM.

For the `DATE` column, we can check the range by looking for the difference in the minimum and maximum dates.

```
range.full <- max(df.atm$DATE) - min(df.atm$DATE)
range.full
```

```
## Time difference of 378 days
```

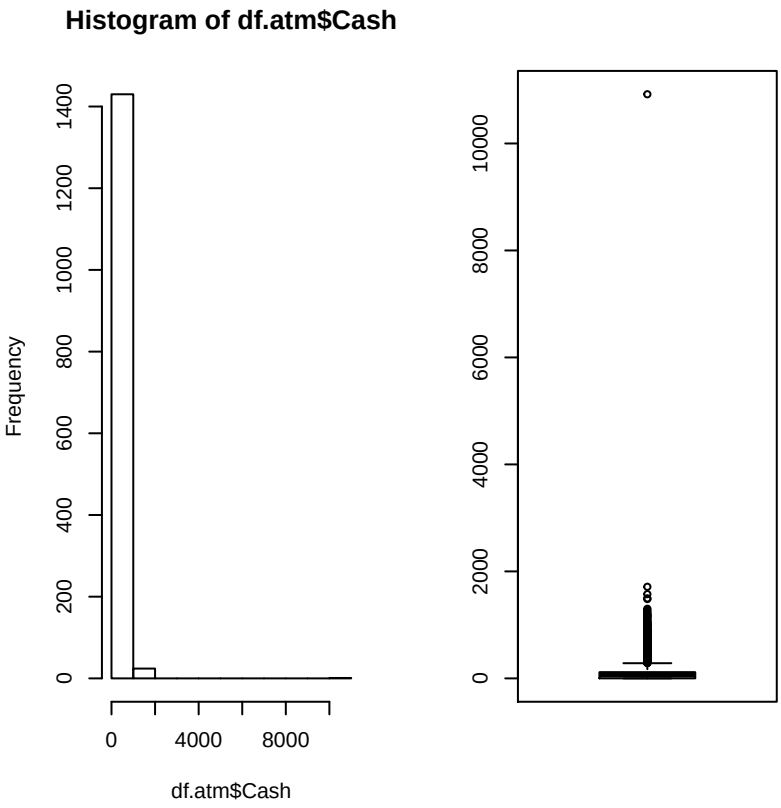
Regarding the `ATM` column, a set of categorical variables, we can check to see the number of unique values.

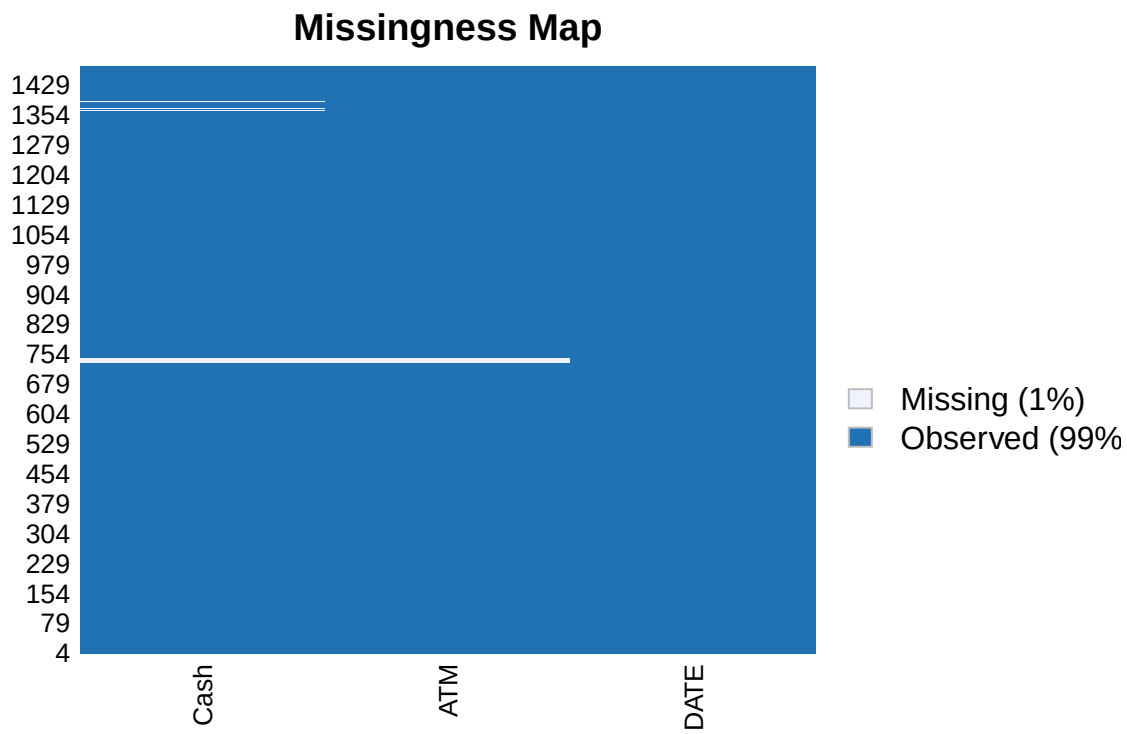
```
unique(df.atm$ATM)
```

```
## [1] "ATM1" "ATM2" NA     "ATM3" "ATM4"
```

As for the `Cash` column, we can create a histogram to understand the distribution of this numeric variable.

```
par(mfrow = c(1,3))
hist(df.atm$Cash)
boxplot(df.atm$Cash)
missmap(df.atm)
```


Histogram of df.atm$Cash
Frequency
0
200
400
600
800
1000
1200
1400
0
4000
8000
df.atm$Cash
0
2000
4000
6000
8000
10000


Missingness Map
1429
1354
1279
1204
1129
1054
979
904
829
754
679
604
529
454
379
304
229
154
79
4
Cash
ATM
DATE
Missing (1%)
Observed (99%

Create Subsets For Each ATM

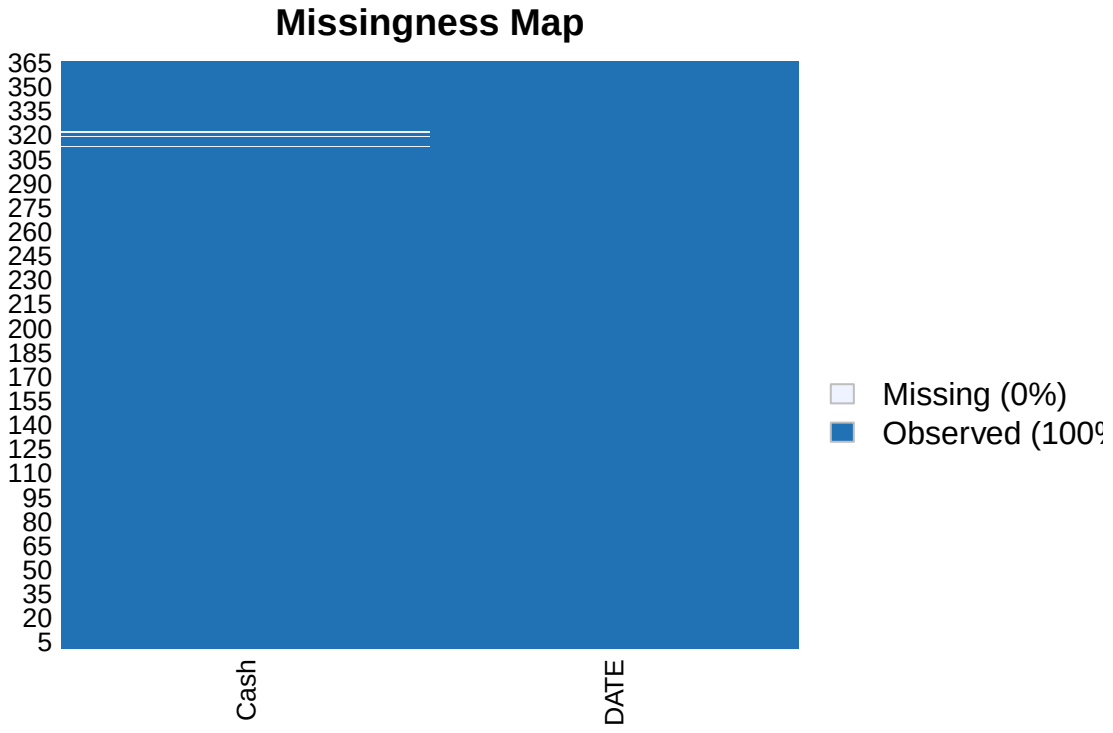
After reviewing the full dataset, we could identify positive right skewness along with some outliers. We'll want to evaluate each of the ATM's subset of the data to eventually create time series objects out of each. Additionally, we'll remove the ATM identifier column as it is no longer needed. From each subset, we can use the `missmap` function to see how many observations are missing, and determine whether we should drop them or try to impute the missing values.

ATM 4

```
df.atm1 <- subset(df.atm, df.atm$ATM == 'ATM1')
df.atm1 <- df.atm1[c('DATE', 'Cash')]
summary(df.atm1)
```

```
##       DATE                 Cash        
##  Min.   :2009-05-01   Min.   :  1.00  
##  1st Qu.:2009-07-31   1st Qu.: 73.00  
##  Median :2009-10-30   Median : 91.00  
##  Mean   :2009-10-30   Mean   : 83.89  
##  3rd Qu.:2010-01-29   3rd Qu.:108.00  
##  Max.   :2010-04-30   Max.   :180.00  
##                       NA's   :3
```

```
missmap(df.atm1)
```

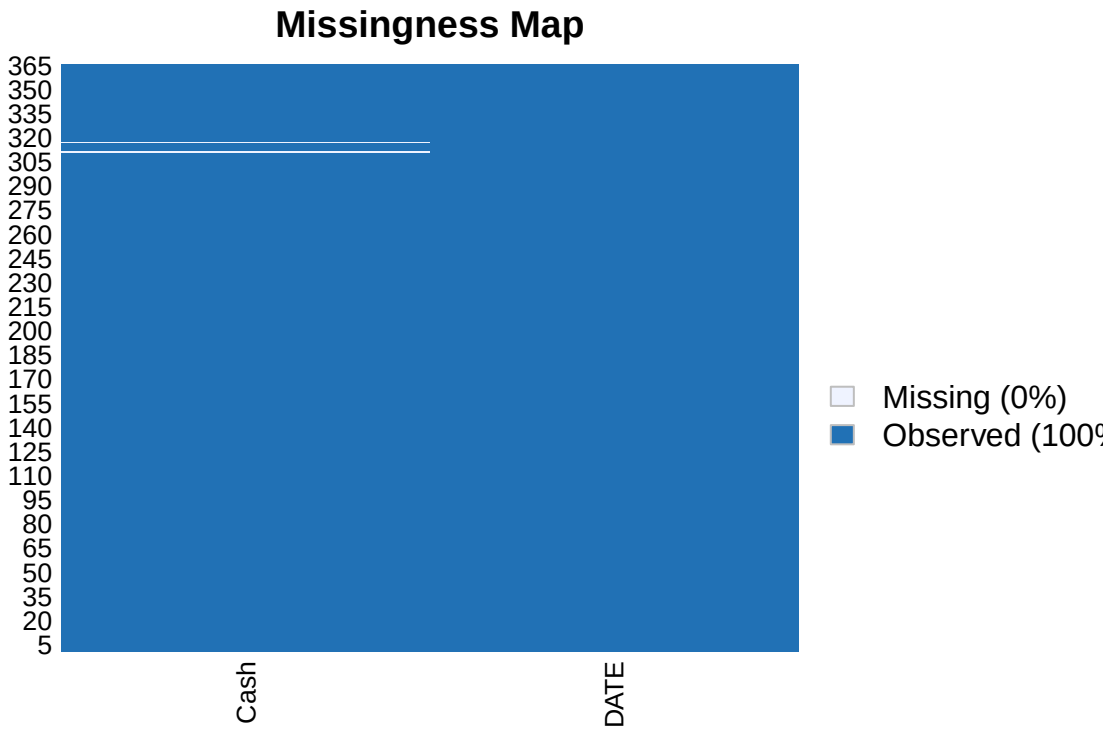


ATM 2

```
df.atm2 <- subset(df.atm, df.atm$ATM == 'ATM2')
df.atm2 <- df.atm2[c('DATE', 'Cash')]
summary(df.atm2)
```

```
##       DATE                 Cash        
##  Min.   :2009-05-01   Min.   :  0.00  
##  1st Qu.:2009-07-31   1st Qu.: 25.50  
##  Median :2009-10-30   Median : 67.00  
##  Mean   :2009-10-30   Mean   : 62.58  
##  3rd Qu.:2010-01-29   3rd Qu.: 93.00  
##  Max.   :2010-04-30   Max.   :147.00  
##                       NA's   :2
```

```
missmap(df.atm2)
```



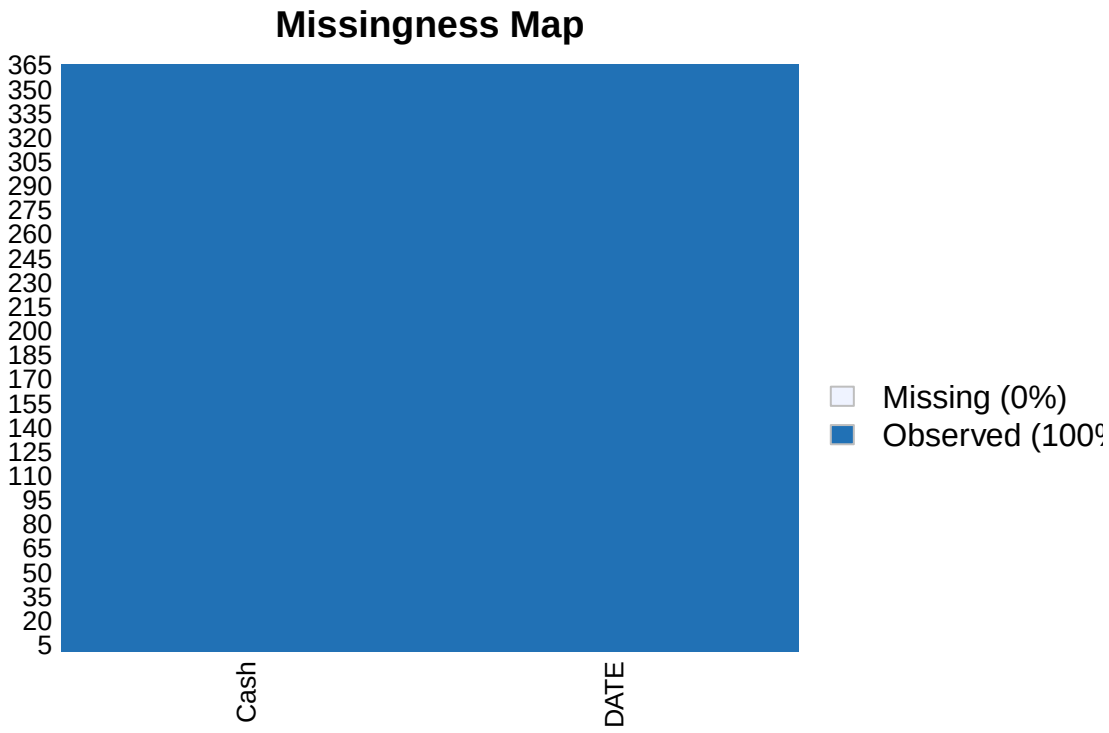
ATM 3

```
df.atm3 <- subset(df.atm, df.atm$ATM == 'ATM3')
df.atm3 <- df.atm3[c('DATE', 'Cash')]
summary(df.atm3)
```

```
##       DATE                 Cash
```

```
##  Min.   :2009-05-01   Min.   : 0.0000
##  1st Qu.:2009-07-31   1st Qu.: 0.0000
##  Median :2009-10-30   Median : 0.0000
##  Mean   :2009-10-30   Mean   : 0.7206
##  3rd Qu.:2010-01-29   3rd Qu.: 0.0000
##  Max.   :2010-04-30   Max.   :96.0000
```

```
missmap(df.atm3)
```

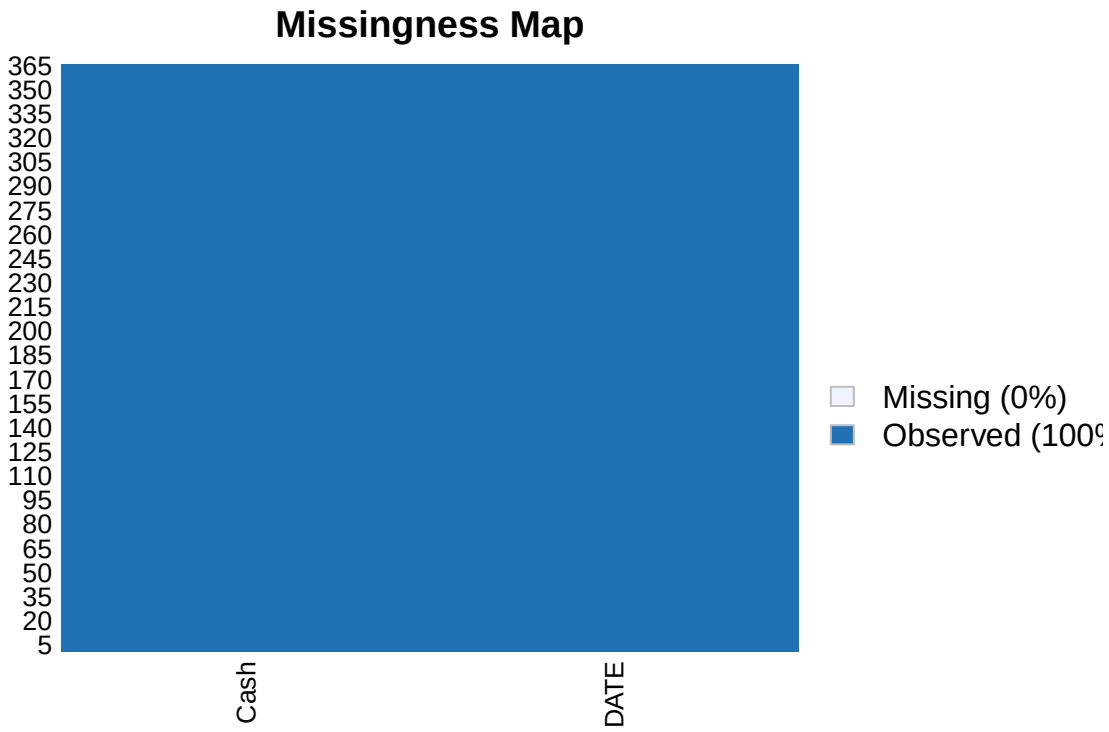


ATM 4

```
df.atm4 <- subset(df.atm, df.atm$ATM == 'ATM4')
df.atm4 <- df.atm4[c('DATE', 'Cash')]
summary(df.atm4)
```

```
##       DATE                 Cash
##  Min.   :2009-05-01   Min.   :    1.563
##  1st Qu.:2009-07-31   1st Qu.:  124.334
##  Median :2009-10-30   Median :  403.839
##  Mean   :2009-10-30   Mean   :  474.043
##  3rd Qu.:2010-01-29   3rd Qu.:  704.507
##  Max.   :2010-04-30   Max.   :10919.762
```

```
missmap(df.atm4)
```



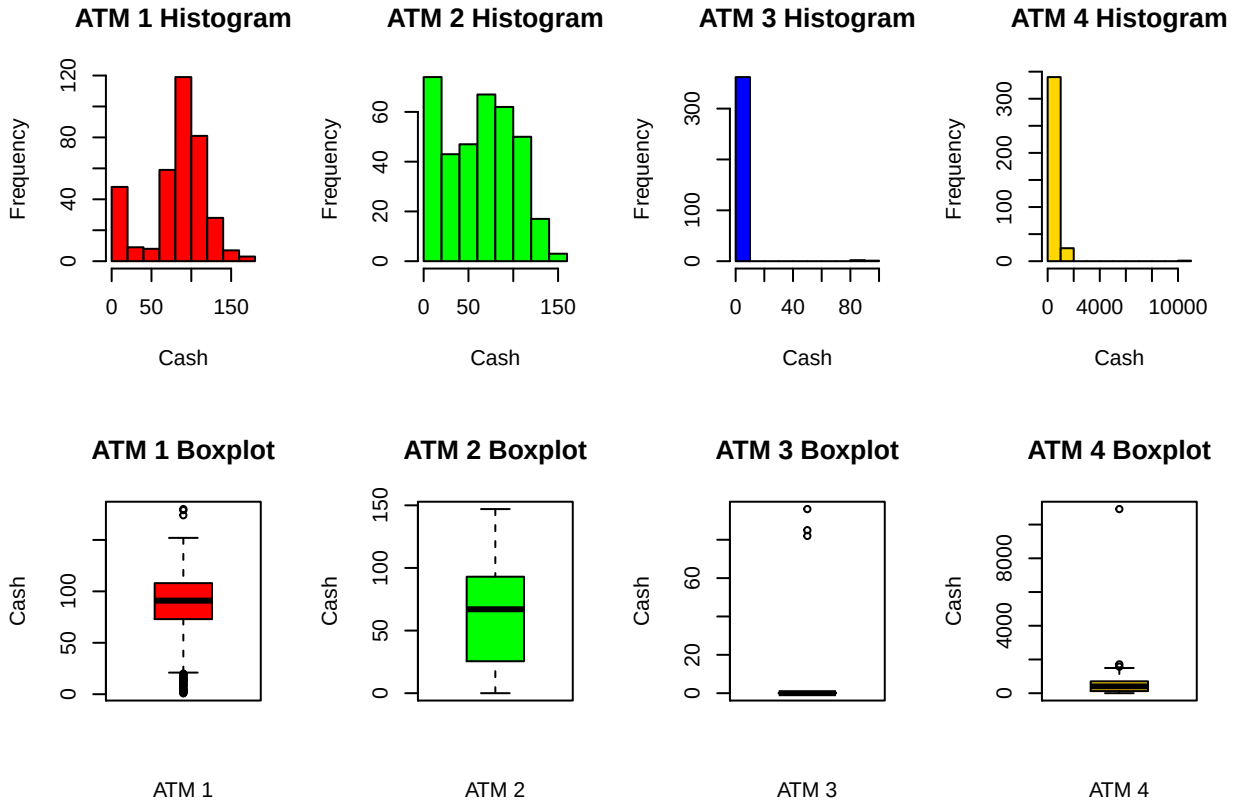
We can see that ATM's 1 and 2 were the only ATM's with missing values, 3 and 2 respectively. With this minimal amount of missing values, we'll drop them from the two subsets.

```
df.atm1 <- drop_na(df.atm1)
df.atm2 <- drop_na(df.atm2)
```

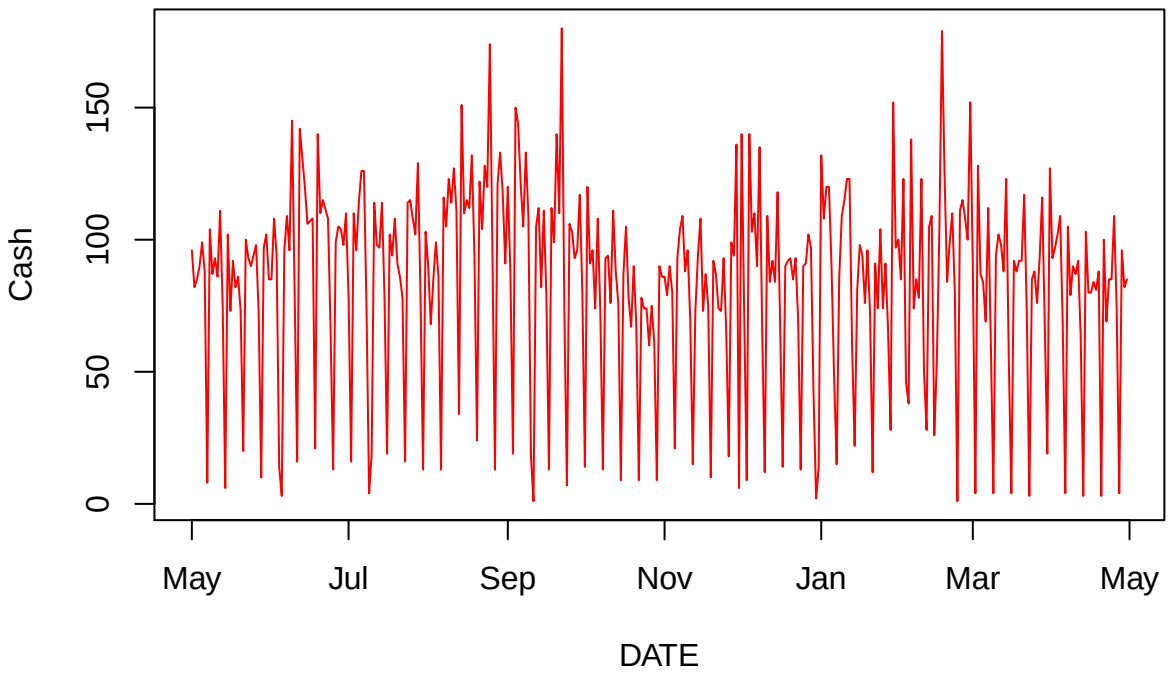
Comparison of Distributions For Each ATM

With the subsets created, missing values removed, we can evaluate each of the respective subset distributions along with the cash withdrawals overtime.

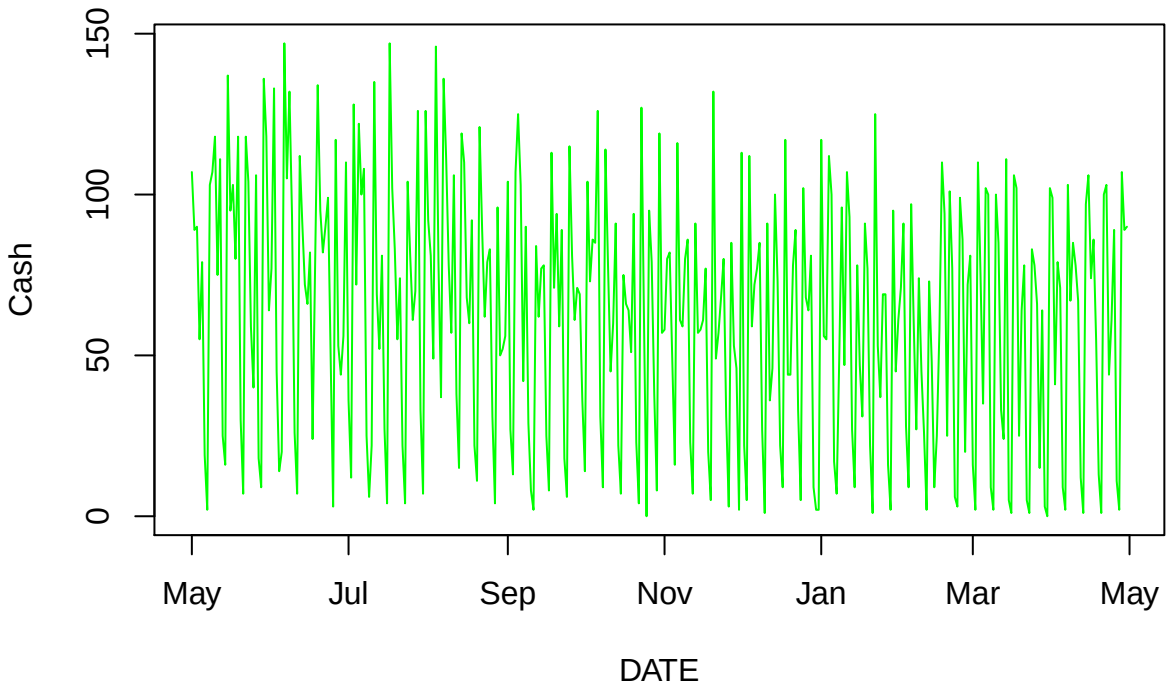
```
par(mfrow = c(2,4))
hist(df.atm1$Cash, main = 'ATM 1 Histogram', xlab = 'Cash', ylab = 'Frequency', col = 'red')
hist(df.atm2$Cash, main = 'ATM 2 Histogram', xlab = 'Cash', ylab = 'Frequency', col = 'green')
hist(df.atm3$Cash, main = 'ATM 3 Histogram', xlab = 'Cash', ylab = 'Frequency', col = 'blue')
hist(df.atm4$Cash, main = 'ATM 4 Histogram', xlab = 'Cash', ylab = 'Frequency', col = 'gold')
boxplot(df.atm1$Cash, main = 'ATM 1 Boxplot', xlab = 'ATM 1', ylab = 'Cash', col = 'red')
boxplot(df.atm2$Cash, main = 'ATM 2 Boxplot', xlab = 'ATM 2', ylab = 'Cash', col = 'green')
boxplot(df.atm3$Cash, main = 'ATM 3 Boxplot', xlab = 'ATM 3', ylab = 'Cash', col = 'blue')
boxplot(df.atm4$Cash, main = 'ATM 4 Boxplot', xlab = 'ATM 4', ylab = 'Cash', col = 'gold')
```



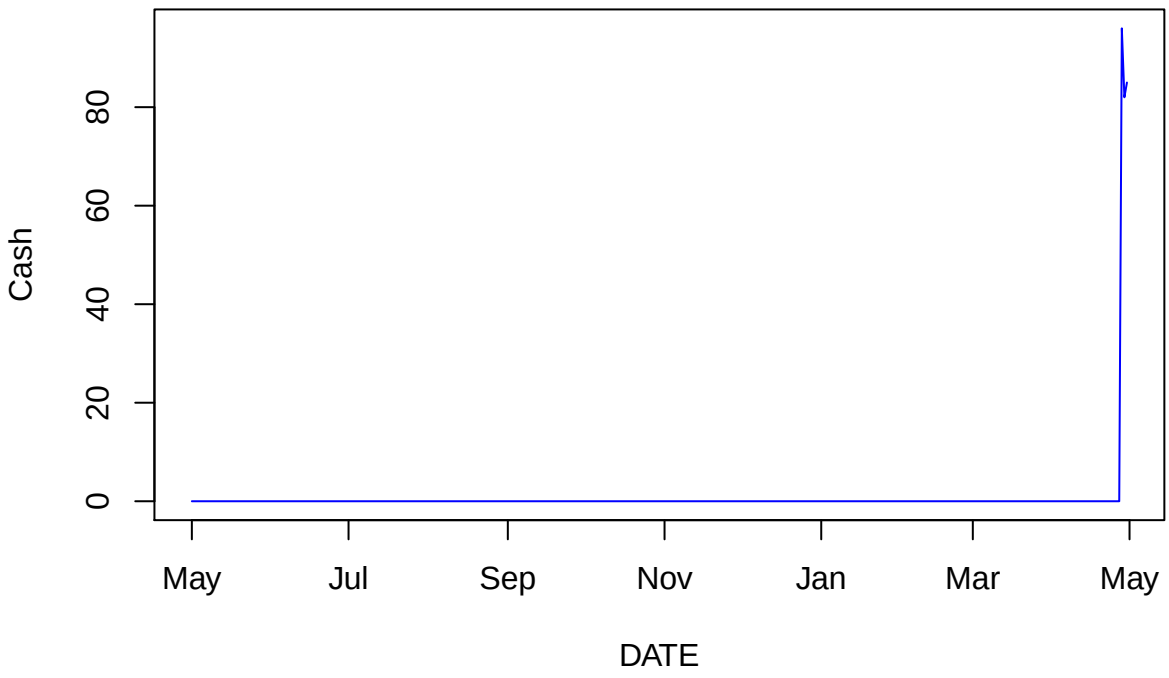

```
plot(df.atm1, type = 'l', col = 'red')
```



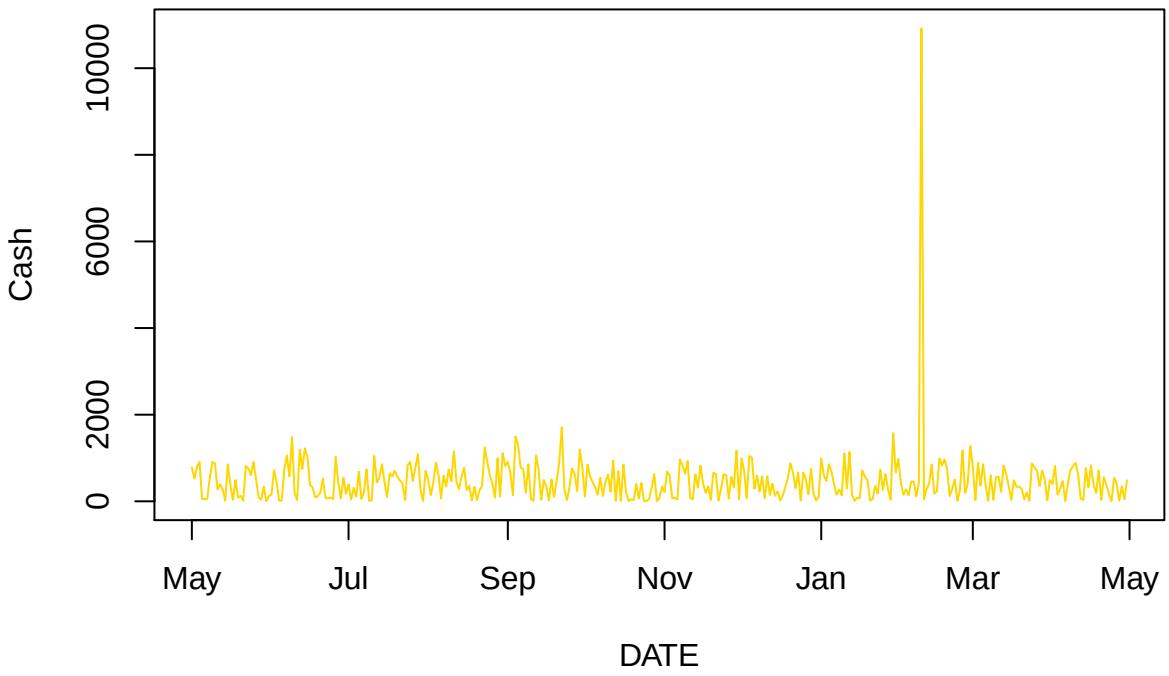

```
plot(df.atm2, type = 'l', col = 'green')
```




```
plot(df.atm3, type = 'l', col = 'blue')
```




```
plot(df.atm4, type = 'l', col = 'gold')
```

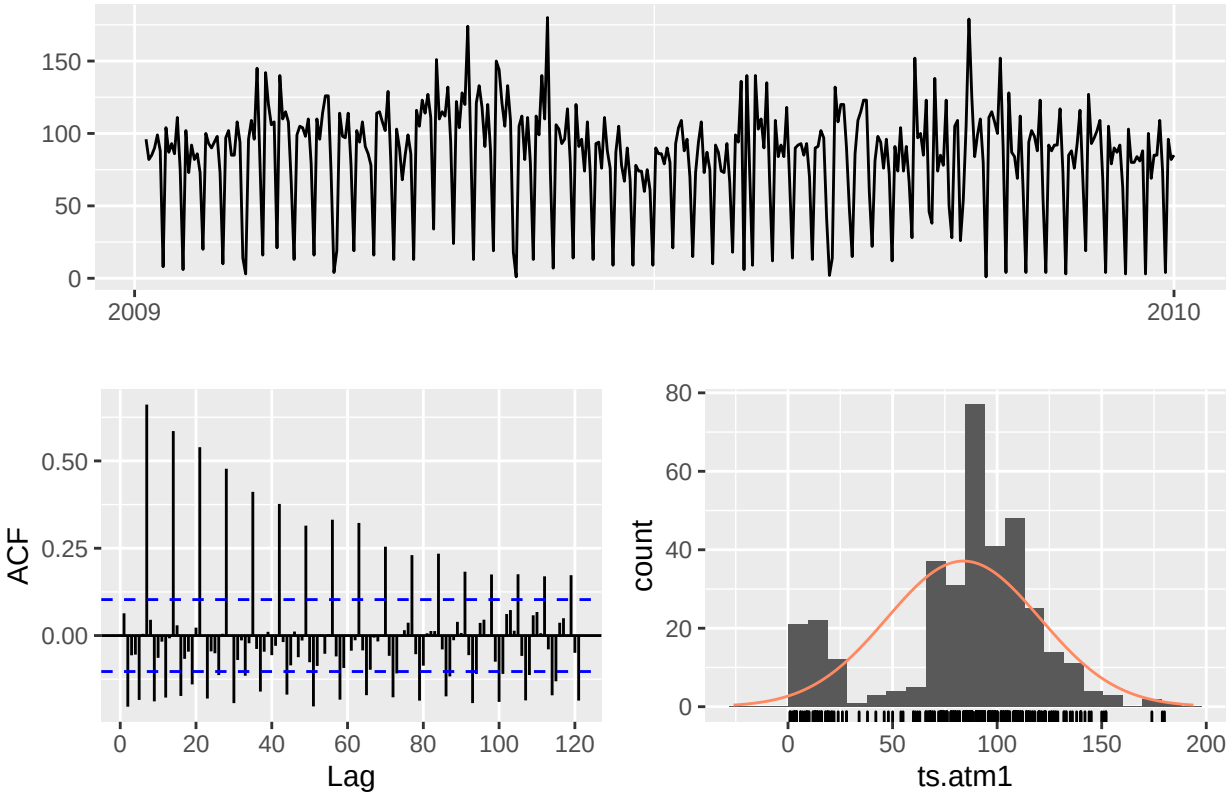

Cash
0
2000
6000
10000
May
Jul
Sep
Nov
Jan
Mar
May
DATE

Time Series Analysis

To get started, we'll need to take our subsets and create a time series object for each ATM using the `ts()` function. We'll specify that the start date for each time series is May 1st, 2009, and the frequency is daily.

```
ts.atm1 <- ts(df.atm1$Cash, start = c(2009, 5, 1), frequency = 365)
ts.atm2 <- ts(df.atm2$Cash, start = c(2009, 5, 1), frequency = 365)
ts.atm3 <- ts(df.atm3$Cash, start = c(2009, 5, 1), frequency = 365)
ts.atm4 <- ts(df.atm4$Cash, start = c(2009, 5, 1), frequency = 365)
```

```
ggtsdisplay(ts.atm1, points = FALSE, plot.type = "histogram")
```



Part B - Forecasting Power