DATA 624 Project 1

Susanna Wong
Professor Joseph Sabeljia
October 25, 2025

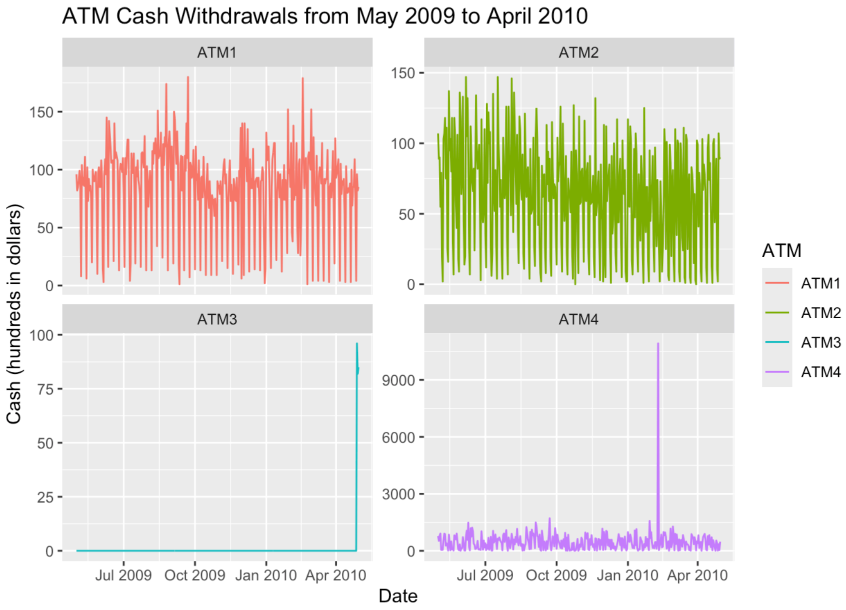
Introduction

This project analyzes three time series dataset - ATM cash withdrawals, residential power consumption, and water flow for two pipes. The goal is to make a forecast on all three datasets.

ATM Cash Withdrawal

The goal is to forecast ATM cash withdrawal from each ATM machines for May 2010. Accurate ATM withdrawal forecasts help optimize cash logistics.

This time series contain 1474 daily observations of 4 different ATM cash machine withdrawal from May 2009 to May 2010. Exploratory analysis reveals seasonal patterns, missing data, and outliers.



Missing Data

There are 19 observations with missing data in the amount of cash that was withdrawn. 14 of these missing values occur in May 2010, the month we aim to forecast. We remove these observations from the training set since we will be forecasting them. The remaining 5 observations are missing cash values in June

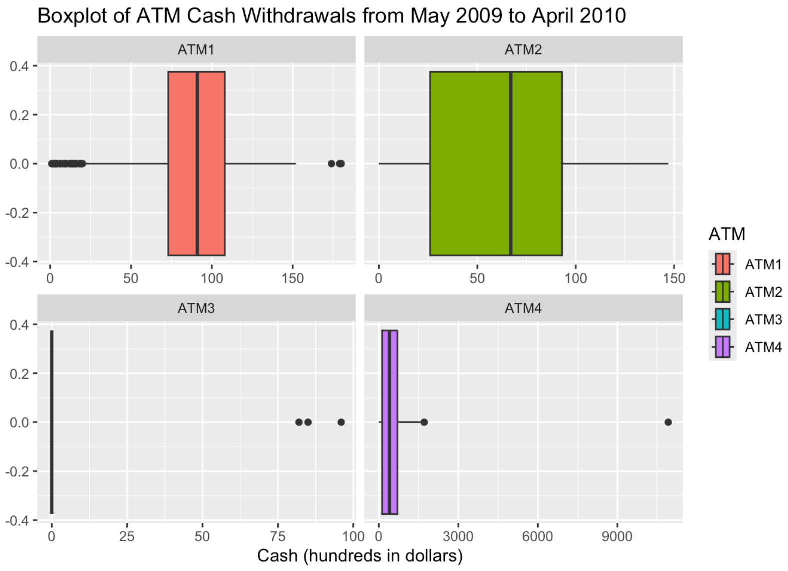
2009, with three from ATM1 and two from ATM2. Using `na.interp()` to fill the missing cash values will help preserves the trend & seasonality. Simply replacing the missing value with the mean, median, or $0 cash value of the relevant ATM will ignore the trends and seasonality.

Outliers

Major outliers are present in ATM 3 and ATM4.

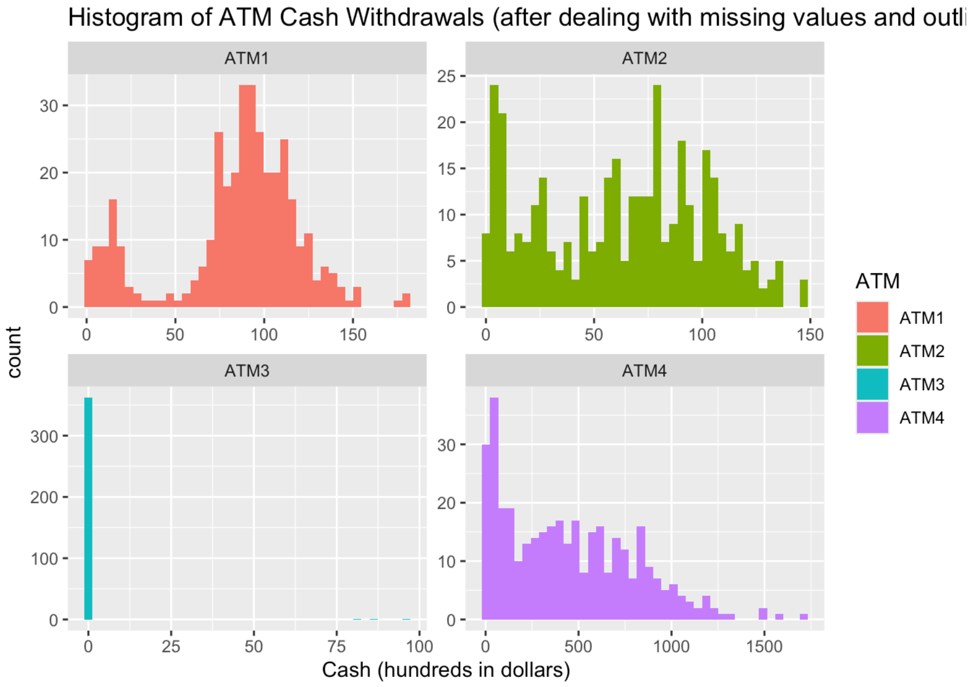
ATM3's values are zeros for almost the entire time until late April 2010. Only three days contain nonzero cash values. There could be various explanation includes that this a fairly new machine so there are no prior transactions recorded on the machine, or perhaps the daily withdrawals were too small that it was rounded down to 0 when rounded to hundreds. Note all cash values in this dataset were integers. So, they could all be rounded values. Regardless, forecasting this ATM will not be meaningful as there are no clear pattern from the 3 days that contain nonzero cash values. We can either exclude this ATM or forecast a simple naive model.

ATM 4 has a large outlier on February 9th, 2010 with an unusual cash withdrawal of $1,092,000. Online research reveals there were a major snowstorm in eastern United States that caused flights cancellation and power outages and XXI Olympic Winter Games occurring in Vancouver, Canada around that time. Even so, there is no available information from the dataset that connects the February 9th spike to the events above. Using `tsoutliers()`to impute the outliers can help maintain the trend and seasonality while imputing the outliers.



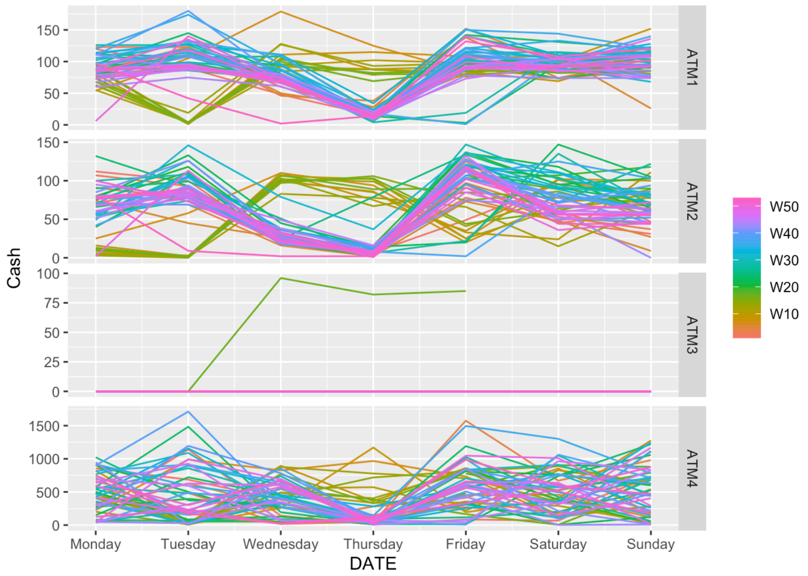
Distribution

ATM1 and ATM2 cash values appear approximately normal. ATM3 is heavily skewed right due to contain almost entirely zeros. ATM4 is right skewed. A box cox transformation was applied to stabilize variance before modeling.



Seasonality and Autocorrection

Visual inspections of seasonal, ACF, and PACF plots reveals strong weekly seasonality. Cash withdrawals seem to peak on Tuesday and Friday for ATM1 and ATM2. For all ATM's except ATM3, ACF and PACF reveals lags in 7, 14, and 21 which is important in modeling choices.



Modeling

Kpss test reveals ATM1 and ATM2 are non- stationary while the others were stationary. One seasonal difference is required to make ATM1 and ATM2 stationary. No further first order difference is needed. The ACF plot shows all autocorrections are within the threshold limit and the residuals are behaving like white noises.

ATM	Best Model	AIC
ATM1	ARIMA(0,0,1)(0,1,2) [7]	3288
ATM2	ARIMA(2,0,2)(0,1,1)[7]	3319
ATM3	Naive	

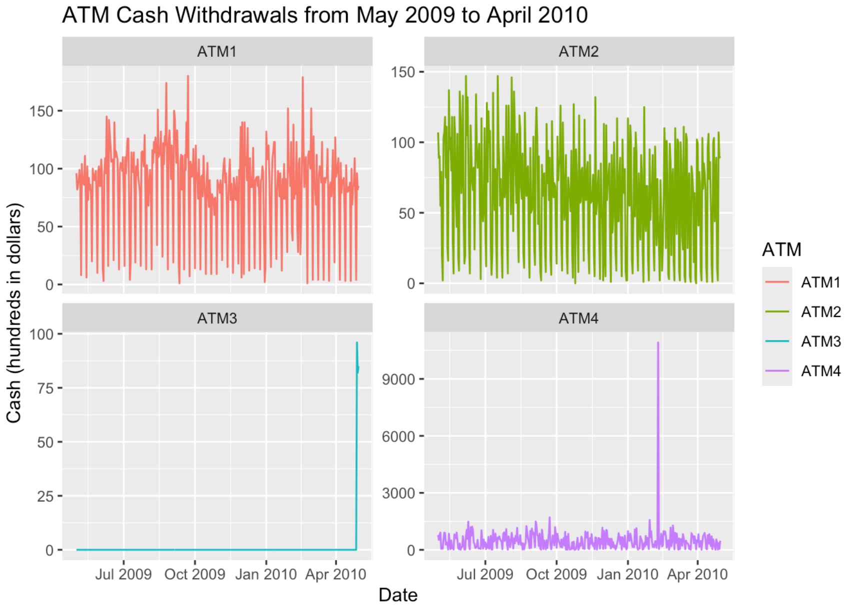
ATM	Best Model	AIC
ATM4	ARIMA(0,0,1)(2,0,0)[7]	2930.7

Residential Power Consumption

The goal is to perform a monthly forecast residential power consumption for for 2014.

This time series contain 192 monthly observations of power consumption from January 1998 to December 2013. Exploratory analysis reveals seasonal patterns, missing data, and outliers.

Exploratory analysis reveals seasonal patterns, missing data, and outliers.



Missing Data

There is only one observations with missing data in the amount of power used in September 2008. 14 of these missing values occur in May 2010, the month we aim to forecast. Since residential power consumption is likely seasonal by the day of the week or month, we can use `na.interp()` to fill the value. This will help