Eexploratory Data Analysis

Enqun Wang (EW), Yiyan Zhou (YZ)

May 3, 2016

Input the data

```
setwd("/Users/ewenwang/Dropbox/Data Science/DMAIC/Case Study/3-Analyze")

require(dplyr)

df = data.frame(read.csv("data.csv", header = T)[-c(1:4),-c(2:3)])
```

Preprocess data

```
require(caret)
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

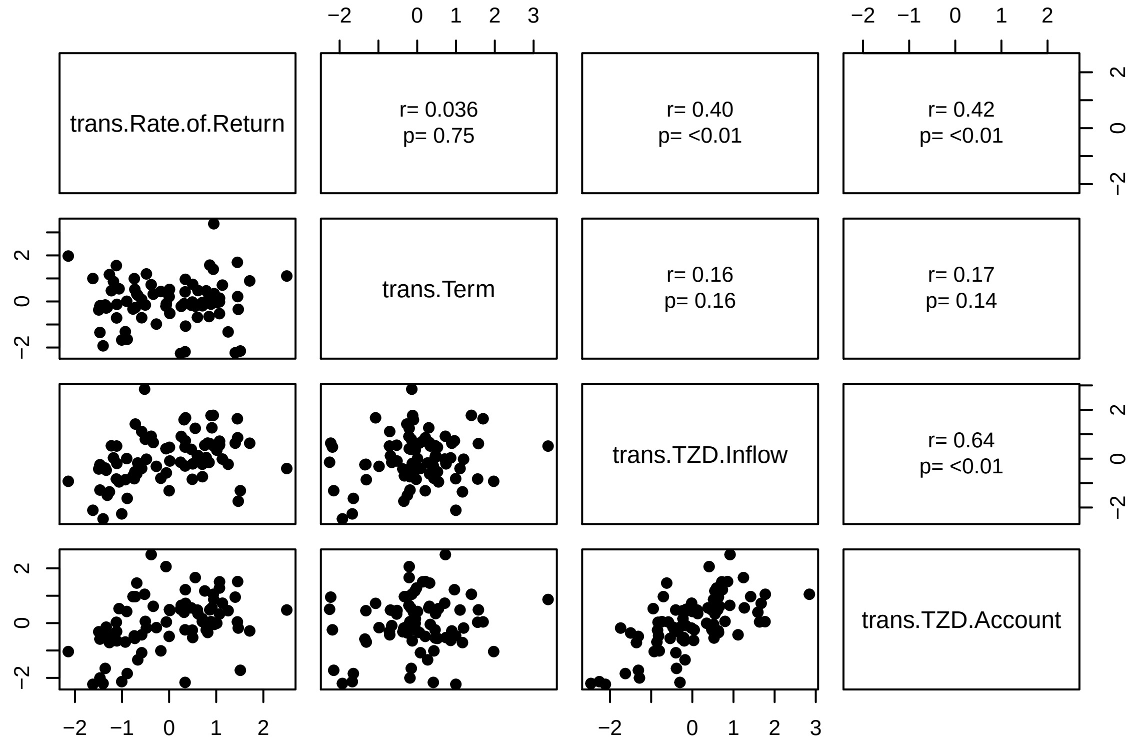
```
date = df[,1]
trans = preProcess(df[,-1], c("BoxCox", "center", "scale"))
predictorsTrans = data.frame(trans = predict(trans, df[,-1]))

X = predictorsTrans[,-1]
y = predictorsTrans[,1]
trans.df = data.frame(data = df[,1], y, X)
```

Based on cause and effect relationship, we divide the variales into four causes.

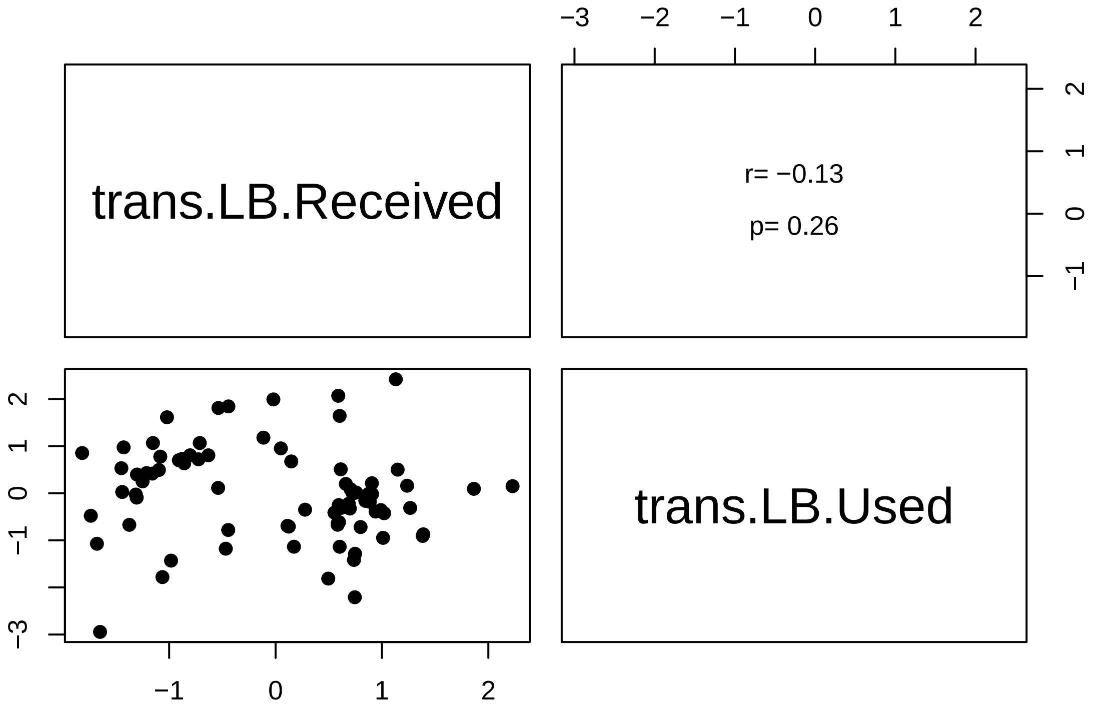
```
X.product = X[, c(1, 2, 5, 6)]
X.promotion = X[, c(7, 8)]
X.platform = X[, c(3, 4)]
X.market = X[, c(9, 10, 11, 12, 13)]
```

Detect dependent variables

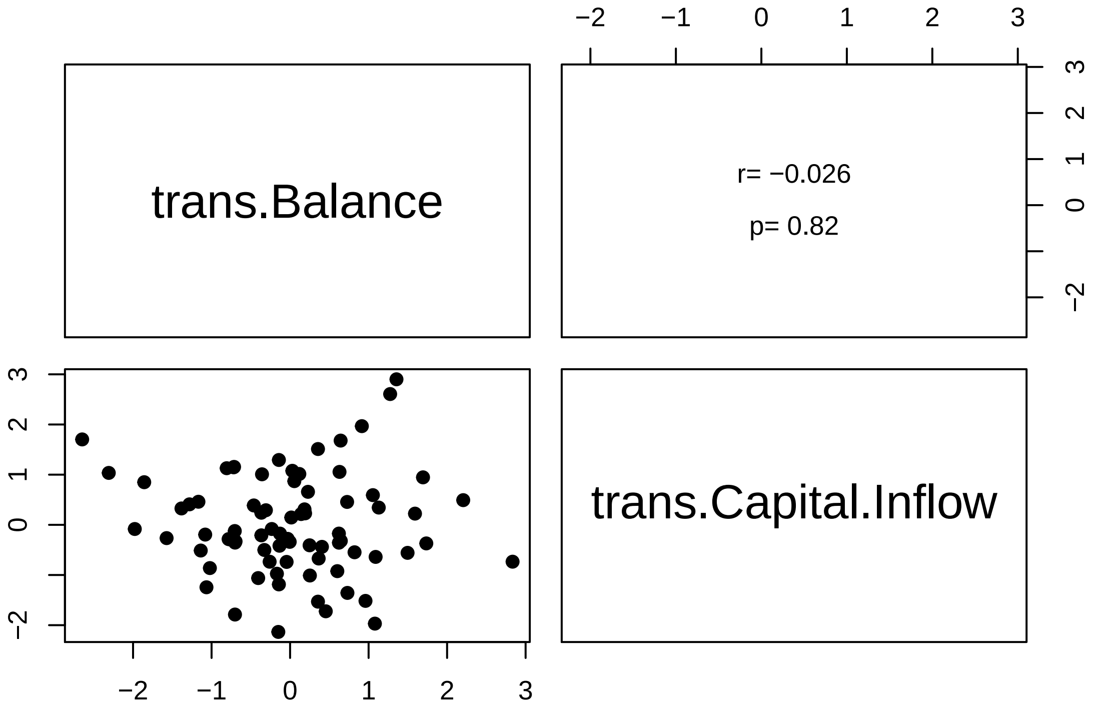
```
panel.cor <- function(x, y, digits = 2, cex.cor, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  # correlation coefficient
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste("r= ", txt, sep = "")
  text(0.5, 0.6, txt)
  # p-value calculation
  p <- cor.test(x, y)$p.value
  txt2 <- format(c(p, 0.123456789), digits = digits)[1]
  txt2 <- paste("p= ", txt2, sep = "")
  if (p < 0.01)
    txt2 <- paste("p= ", "<0.01", sep = "")
  text(0.5, 0.4, txt2)
}

pairs(X.product, pch = 19, upper.panel = panel.cor)
```



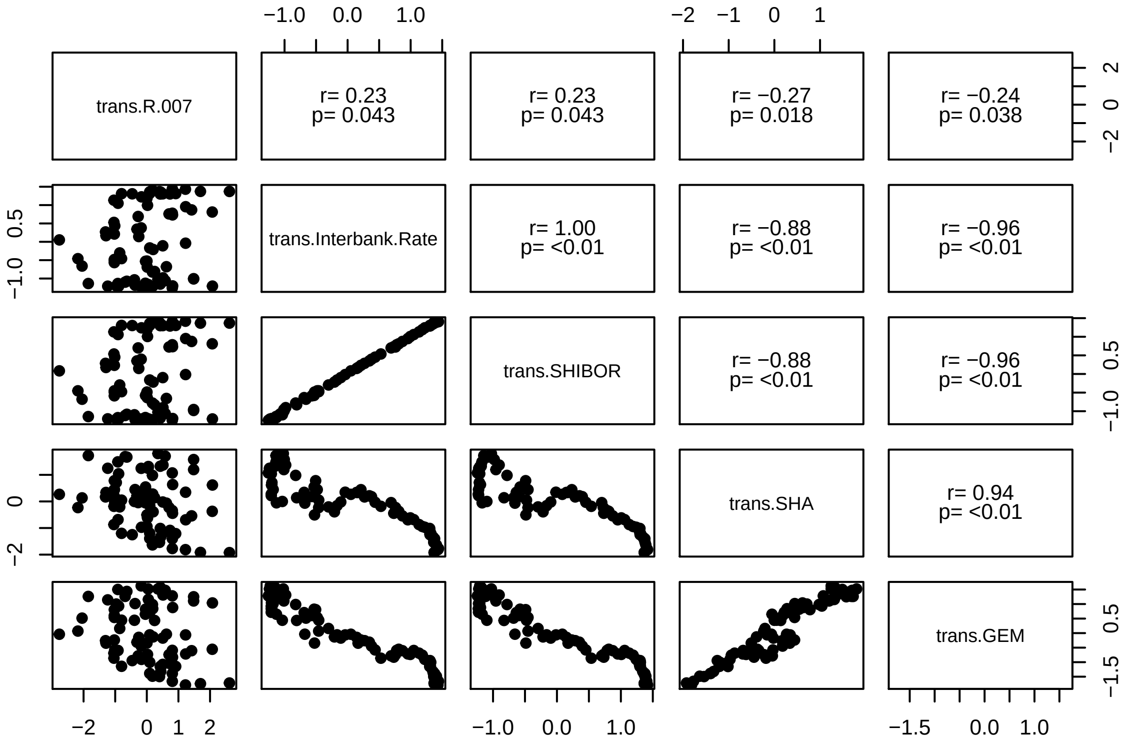
```
pairs(X.promotion, pch = 19, upper.panel = panel.cor)
```



```r
pairs(X.platform, pch = 19, upper.panel = panel.cor)
```



```r
pairs(X.market, pch = 19, upper.panel = panel.cor)
```



According to the correlation plot above, we find that `interbank.Rate` , `SHIBOR`, `SHA`, and `GEM` are highly correlated, and that `TZD.Inflow` and `TZD.Account` are highly correlated. So we consider if we could remove some of them.

Based on the voice of costumers (VOC), we decided to remove `interbank.Rate`, which can be represented by `SHIBOR`; remove `GEM`, which can be represented by `SHA`; and remove `TZD.Inflow`, which can be reflected from `TZD.Account`.

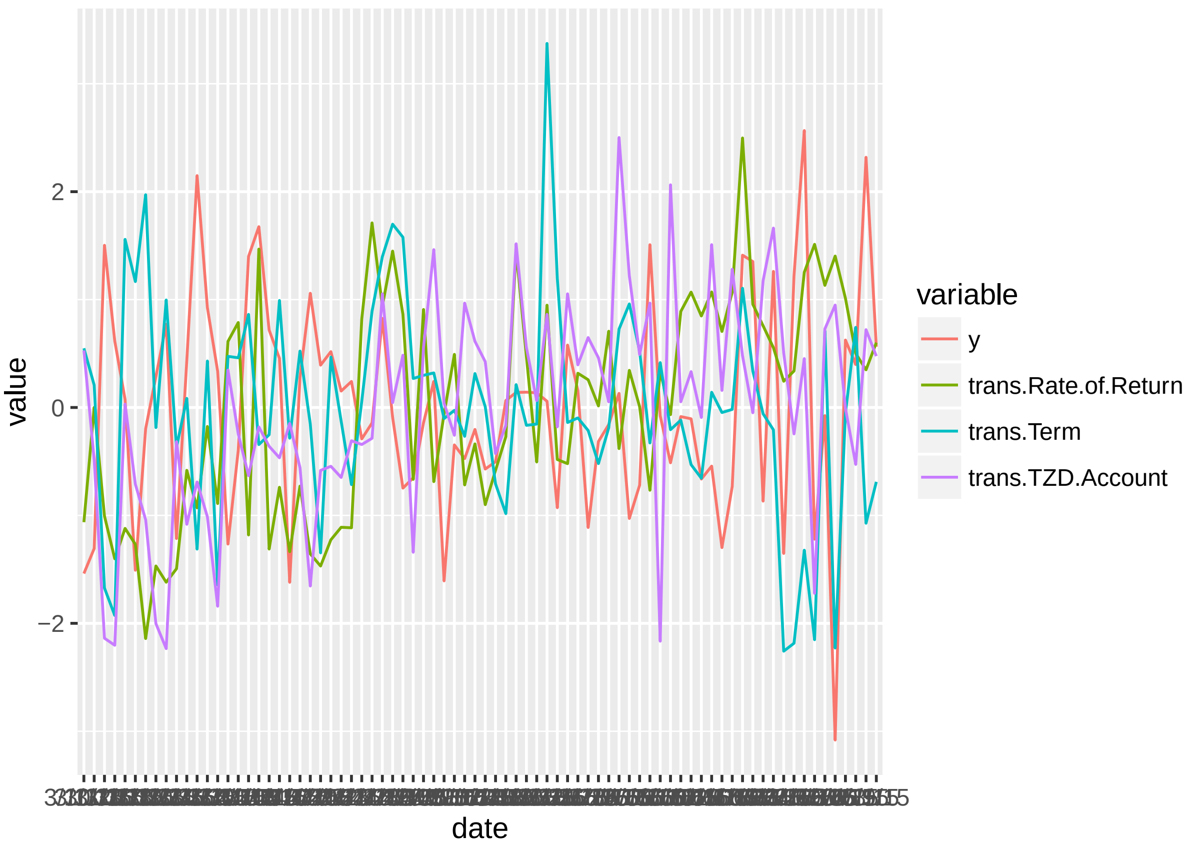
```
X.product = X[, c(1, 2, 6)]
X.market = X[, c(9, 11, 12)]
```

Plot multiple time series

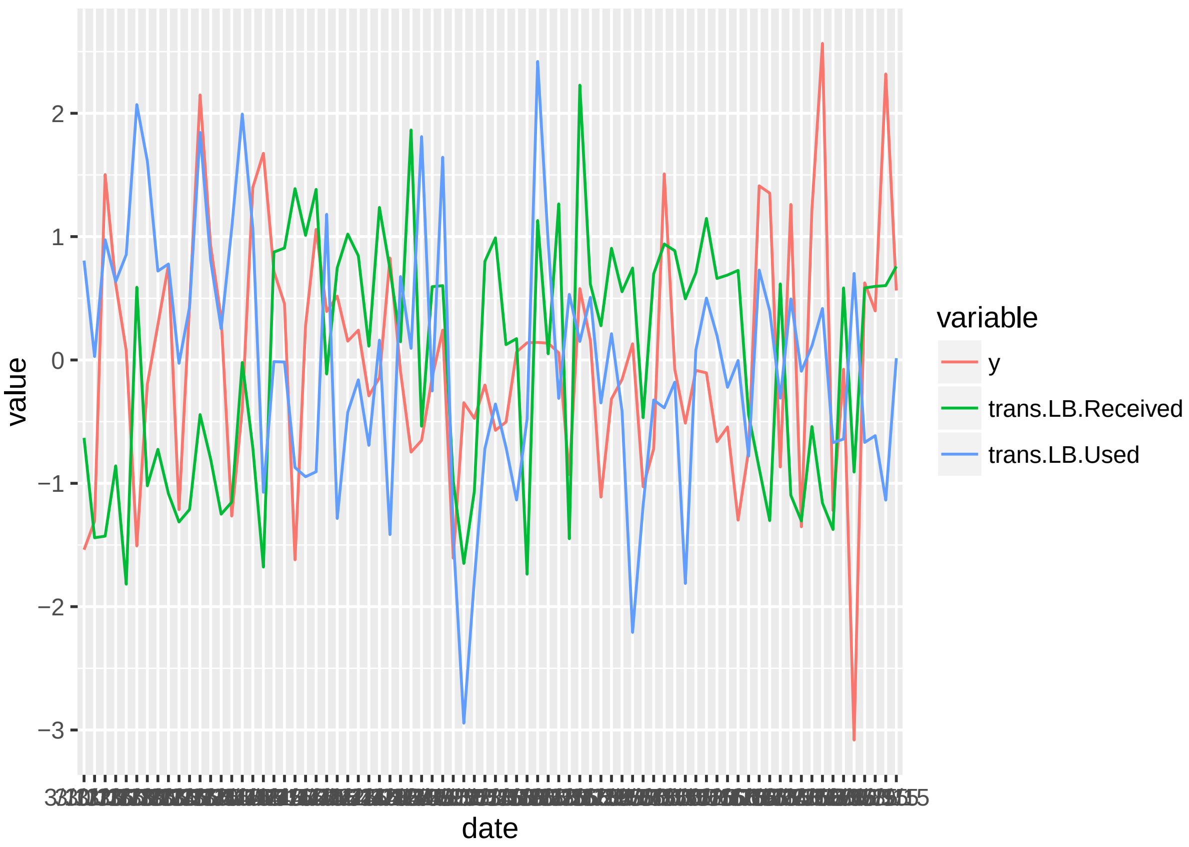
```
require(ggplot2)
require(reshape2)
```

```
## Loading required package: reshape2
```

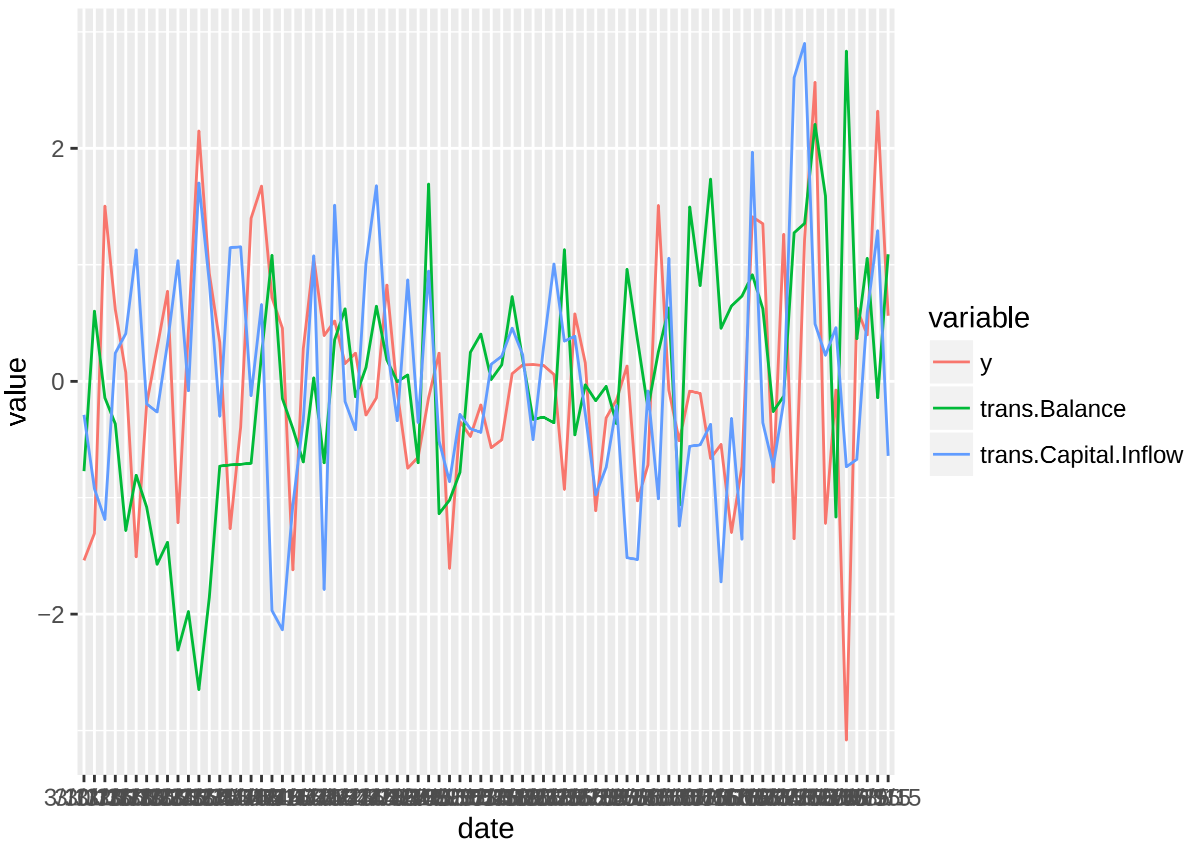
```
df.product = data.frame(date, y, X.product)
meltdf <- melt(df.product, "date")
ggplot(meltdf,aes(x=date,y=value,colour=variable,group=variable)) +
  geom_line()
```



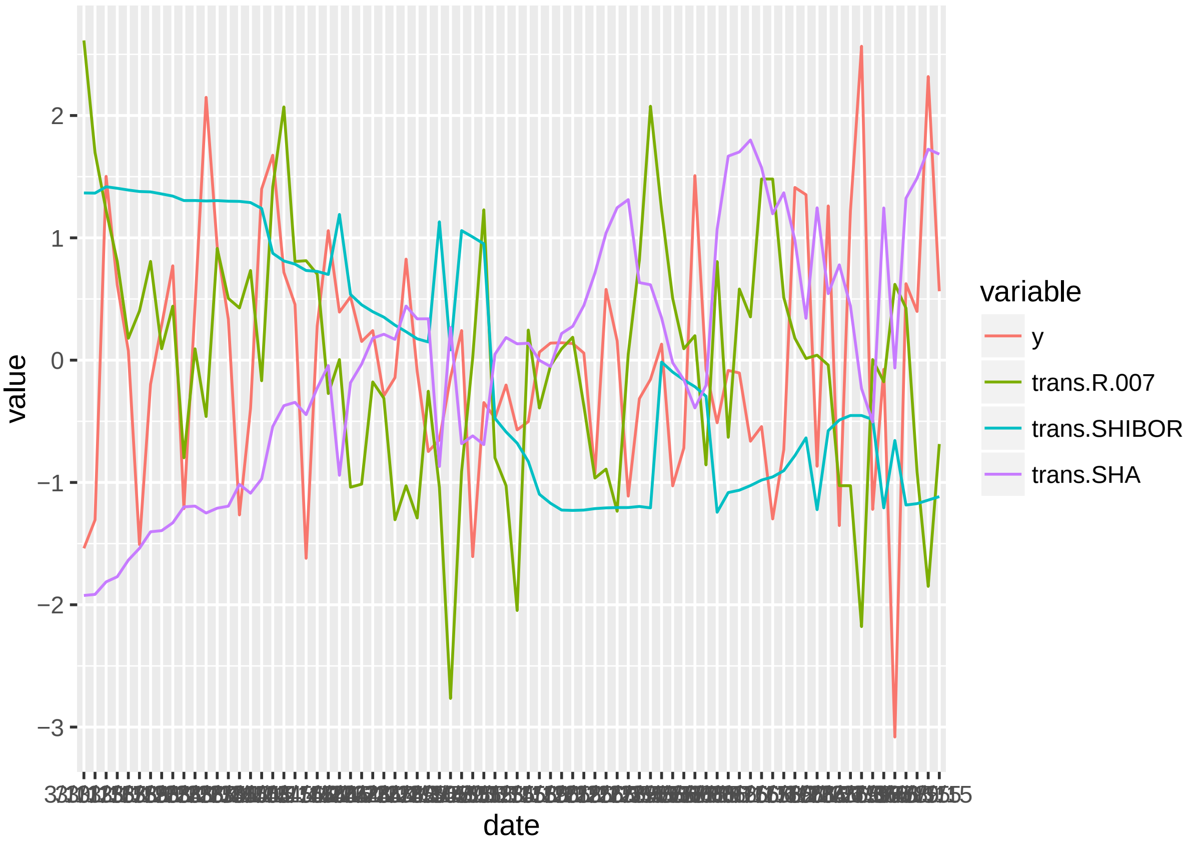
```
df.promotion = data.frame(date, y, X.promotion)
meltdf <- melt(df.promotion, "date")
ggplot(meltdf,aes(x=date,y=value,colour=variable,group=variable)) +
  geom_line()
```



```
df.platform = data.frame(date, y, X.platform)
meltdf <- melt(df.platform , "date")
ggplot(meltdf,aes(x=date,y=value,colour=variable,group=variable)) +
  geom_line()
```



```r
df.market = data.frame(date, y, X.market)
meltdf <- melt(df.market, "date")
ggplot(meltdf,aes(x=date,y=value,colour=variable,group=variable)) +
  geom_line()
```



According to the plots above and VOC, we would remove the variable `SHIBOR` and `SHA`.

First selection of variables

According to the exploratory data analysis, we decide to first elect variables as follows,

- Product Factor:

1. Rate of Return
2. Term
3. TZD Account

- Promotion Factor:

1. LB Received
2. LB Used

- Platform Factor:

1. Balance
2. Capital Inflow

- Market Factor:

1. R.007