UC Santa Barbara, Media Arts and Technology

Exploring late return rate from SPL

Sergio Rodriguez

January 15, 2015

1 Introduction

According to the Seattle Public Library (SPL) policies[1] an item that is checked-out can be returned based on the following schedule:

Item	Loan Period
Books, magazines, pamphlets, CDs, audiocassettes, albums, videocassettes	21 days
DVDs	14 Days

A user can borrow simultaneously up to 50 items and they can be returned to any library location or book drop.

After identifying all items that are returned after their due date, it would be interesting to explore the following aspects:

- Whether there is a specific month of the year with more items returned late.
- What types of items people tend to return late.
- Is there any specific branch with this behaviour?
- What kind of *subjects* are more likely to be returned late.

2 Querying the Database

2.1 Overview

Here we use `MySQL` as client software in order to connect to the SPL database. Furthermore, it may be desireable or even necessary to perform a statistical analysis in a statistical package rather than in the database. The `DBI` package in `R`[2] provides a uniform, client-side interface to different database management systems, such as `MySQL`. The basic model breaks the interface between the client and the server into three main elements:

1. The *driver* facilitates the communication between the `R` session and a particular type of database management system (`MySQL`);
2. the *connection* encapsulates the actual connection (with the aid of the driver) to a particular database management system and carries out the requested queries;
3. and the result which tracks the status of a query, such as the number of rows that have been fetched and whether or not the query has completed[3].

[1] http://www.spl.org/using-the-library/get-started/check-out-and-return
[2] http://www.r-project.org
[3] http://www.stat.berkeley.edu/ nolan/stat133/Fall05/lectures/SQL-R.pdf

2.2 MySQL Queries

```
library(DBI) # Load libraries for connecting R with MySQL
library(RMySQL)

drv = dbDriver("MySQL") # Specify driver

# Below, the user mat259 establishes a connection, called con,
# to the database on the host tango.mat.ucsb.edu. Since the database is password
# protected, the user need to provide a password to gain access to it.

con = dbConnect(drv,
                user="mat259",
                password="V1sual1zat1on",
                host="tango.mat.ucsb.edu",
                port=3306)

# Here is the advantage of using R: the queried table is saved in the variable t1 so we
# can make further analysis and produce plots.
t1 <- dbGetQuery(con,
"SELECT
    MONTH(cout) AS month_out,
    YEAR(cout) AS year_out,
                itemtype,
                count(itemtype) as num_obs,
                avg(timestampdiff(DAY,cout,cin)) AS due_days_avg,
                min(timestampdiff(DAY,cout,cin)) AS due_days_min,
                max(timestampdiff(DAY,cout,cin)) AS due_days_max,
                SUM(CASE
                WHEN timestampdiff(DAY,cout,cin)>= 21
                THEN
                1
                ELSE 0
                END) AS count_late,
                avg(CASE
                WHEN timestampdiff(DAY,cout,cin) >= 21
                THEN
                1
                ELSE 0
                END) AS avg_late
                FROM
                spl2.inraw
                WHERE (DATE(cout) between '2009-01-01' and '2014-12-31')
                GROUP BY year(cout),month(cout),itemtype;")
```

Once we have executed the previous query we validate that `t1` has been correctly populated:

```
head(t1) # Display first 5 obsevations of table 1

##   month_out year_out itemtype num_obs due_days_avg due_days_min
## 1         1     2009    acart       1         6.00            6
## 2         1     2009     acbk  346706        40.06            0
## 3         1     2009    accas    2543        48.13            0
## 4         1     2009     accd  198059        18.68            0
## 5         1     2009  accdrom      30        57.03            0
## 6         1     2009    acdvd  239292        14.15            0
```

```
##   due_days_max count_late avg_late
## 1            6          0   0.0000
## 2         1769     191443   0.5522
## 3          754       1538   0.6048
## 4          958      61958   0.3128
## 5          767         16   0.5333
## 6          765      43884   0.1834
```

3 Analysis

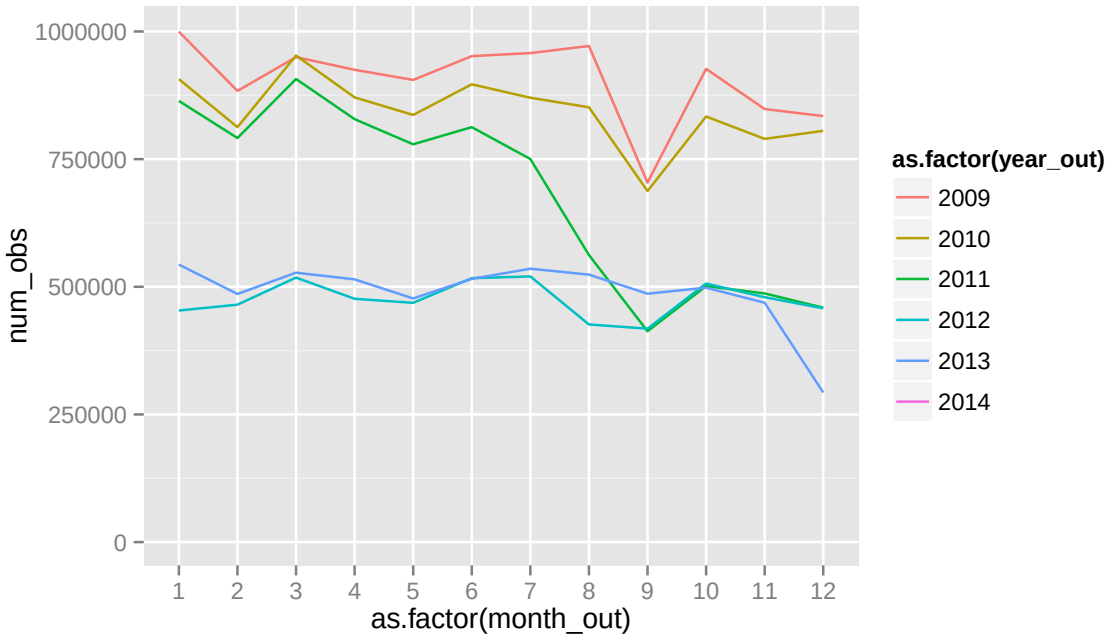
Let us first analyze whether there is a seasonal pattern on the total number of check-outs regardless their type or whether it was returned late or not:

```
options(sqldf.driver = "SQLite")  # For running SQL-like statements locally
options(gsubfn.engine = "R")

library(RMySQL)
library(sqldf)

# Aggregate data per year and month regarless its type
df1  <- sqldf("select sum(num_obs) as num_obs, year_out,month_out
              from 't1'
              group by month_out, year_out
              order by month_out, year_out")

# Plot
library(ggplot2) # load library
p <- ggplot(df1,aes(x=as.factor(month_out),y=num_obs,
                    colour=as.factor(year_out),group=as.factor(year_out)))
p + geom_line()
```

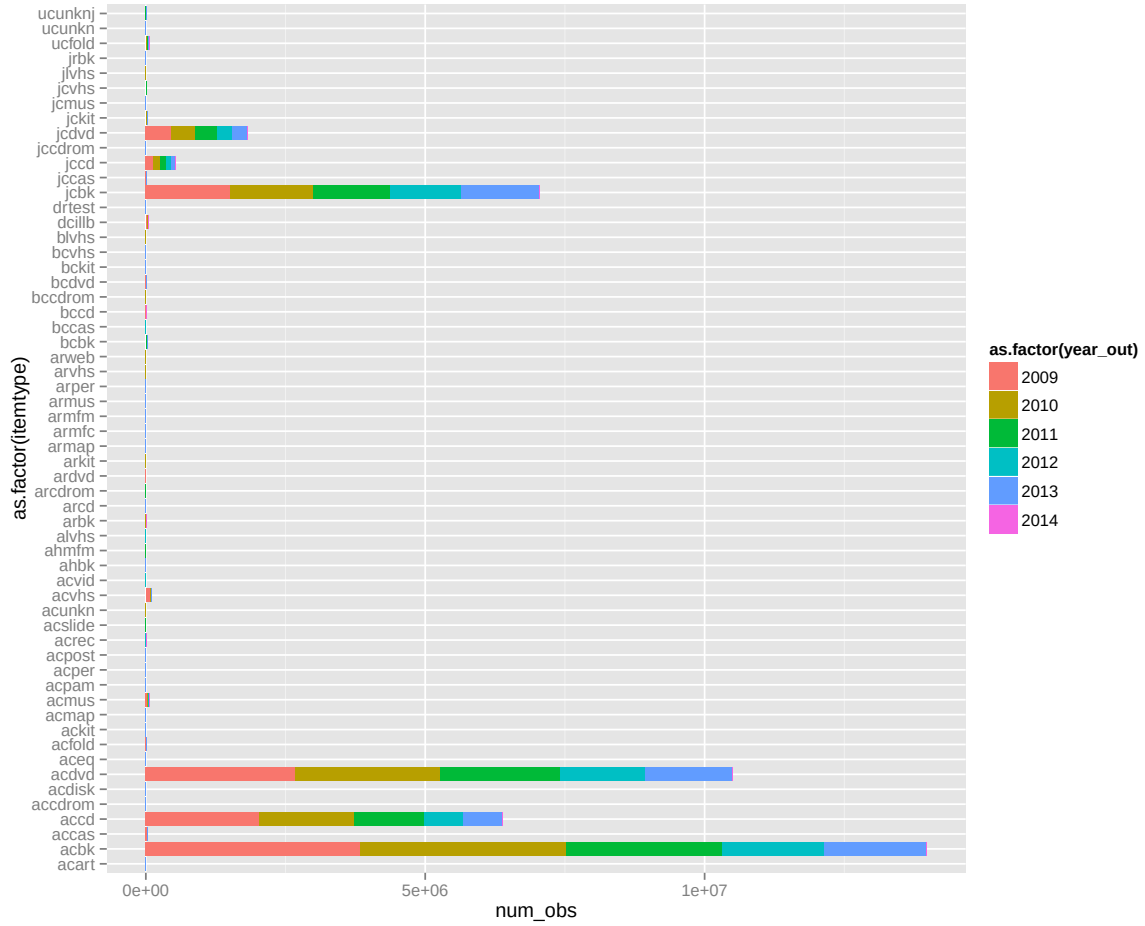


- Overall, there has been a general decreasing in check-outs over time: from 2009 to 2013 check-outs have decreased almost by a half.

- Interestingly, September is the month with less check-outs whereas August and April are the months with highest number of check-outs.

Let us now explore which type of items are the most popular over time:

```
df2  <- sqldf("select sum(num_obs) as num_obs, year_out, itemtype
              from 't1'
              group by year_out,itemtype")
p <- ggplot(df2,aes(x=as.factor(itemtype),y=num_obs,fill=as.factor(year_out)))
p + geom_bar(stat="identity") + coord_flip()
```



From the previous graph we see that the most popular (more checked-out) types are:

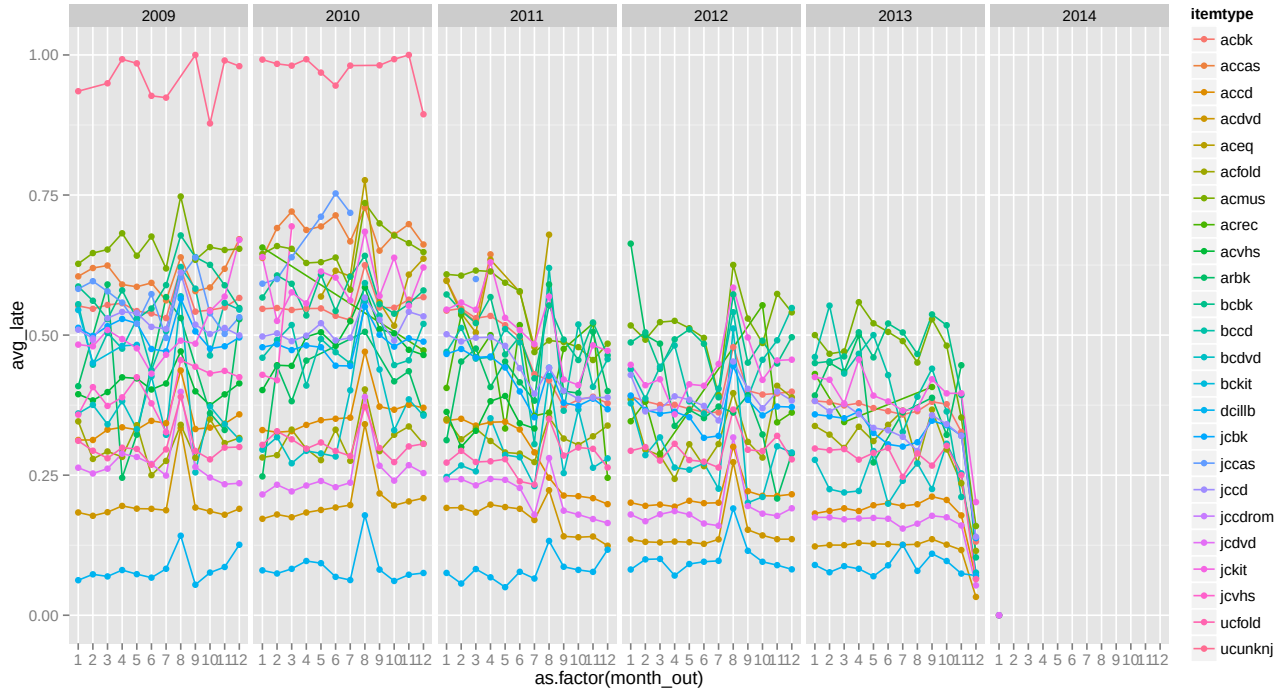
- jcdvd: Media/DVD/Juv Circulating
- jccd: Media/CD/Juvenile
- jcbk: book/Book/Juvenile
- acdvd:media/DVD/Adult/YA
- accd:media/CD/Adult/YA
- acbk:book/Book/Adult

Not surprisingly, the items that have been checked-out the most are DVDs, CDs and books either from the juvenile or adult section. But, back to our original question, are these items the ones that are returned late the most? One would expect that the more items checked-out the more likely to be returned late. Thus, we weight the total number of late returns by total number of check-outs so all types of items are comparable. This is variable

avg_late, which also represents the likelihood of an item to be returned late in a specific month of the year. However, we only turn our attention to the items that are *popular* the most. To do so, we will drop all monthly observations whose value is below the median of the overall yearly demand of items - here is when it becomes handy to work within a statistical analysis software as opposed to a database engine.

```
# Compute yearly median of number of checkouts
df4 <- merge(t1,
      aggregate(num_obs ~ year_out,t1,median),
      by="year_out")
df4 <- subset(df4,num_obs.x>=num_obs.y)
p <- ggplot(df4,
            aes(x=as.factor(month_out),y=avg_late,colour=itemtype,label=itemtype,
                group=itemtype))
p <- p + geom_line()
p <- p + geom_point()

# add faceting
p + facet_grid(. ~ year_out)
```



Notice that August is when the higher rate of late return occurs amongst all items types and restricted only to those items that are checked-out the most.

Finally, we state our main conclusion: the types of items that are checked the most with average late return higher that 75% are the following: ucunknj, jccas and aceq. As seen on the previous plot, this pattern occur consistently over all months of the 5 years analyzed.