Birthdays Analysis

This file details the steps taken to analyze the Birthdays. The file used is a sample file from Udacity 651 course. https://s3.amazonaws.com/udacity-hosted-downloads/ud651/birthdaysExample.csv

```
bd <- read.csv("birthdaysExample.csv")
```

A peek into the dataframe:

```
head(bd)
```

```
##      dates
## 1 11/25/14
## 2   6/8/14
## 3  9/12/14
## 4  5/26/14
## 5  2/20/14
## 6  6/19/14
```

We observe that this dataset has only one column called dates. Now converting this column into POSIX date. Using Lubridate package.

```
require(lubridate)
```

```
## Loading required package: lubridate
```

```
## Warning: package 'lubridate' was built under R version 3.2.3
```

```
dates <- as.data.frame(parse_date_time(bd$dates,"m!*/d!/y!*"))

names(dates) <- c("date")
```

Extract the month and days from the date and add it to the dataframe.

```
require(dplyr)
```

```
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dates <- dates %>% mutate(month = month(date), day = day(date))

head(dates)
```

```
##         date month day
## 1 2014-11-25    11  25
## 2 2014-06-08     6   8
## 3 2014-09-12     9  12
## 4 2014-05-26     5  26
## 5 2014-02-20     2  20
## 6 2014-06-19     6  19
```

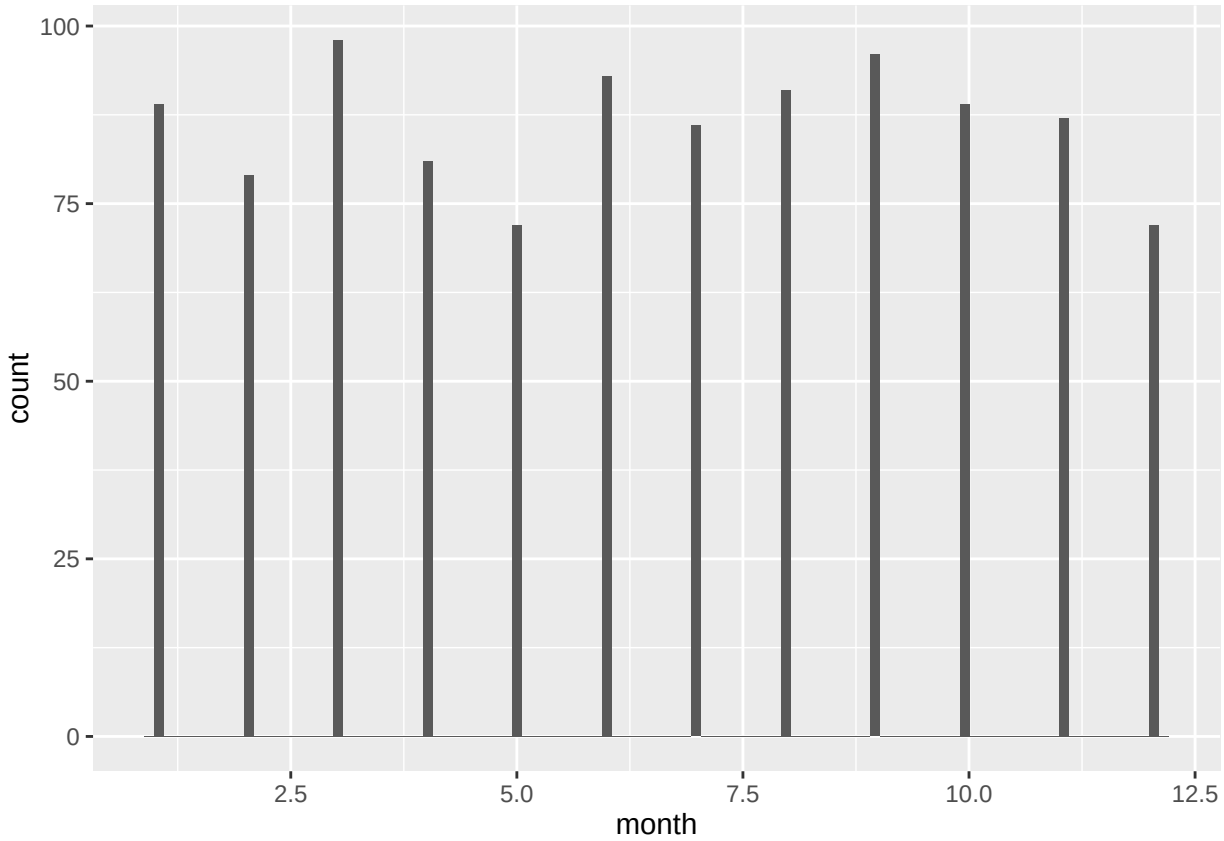
Analyzing the distribution of birthdays across different months:

```
require(ggplot2)
```

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

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

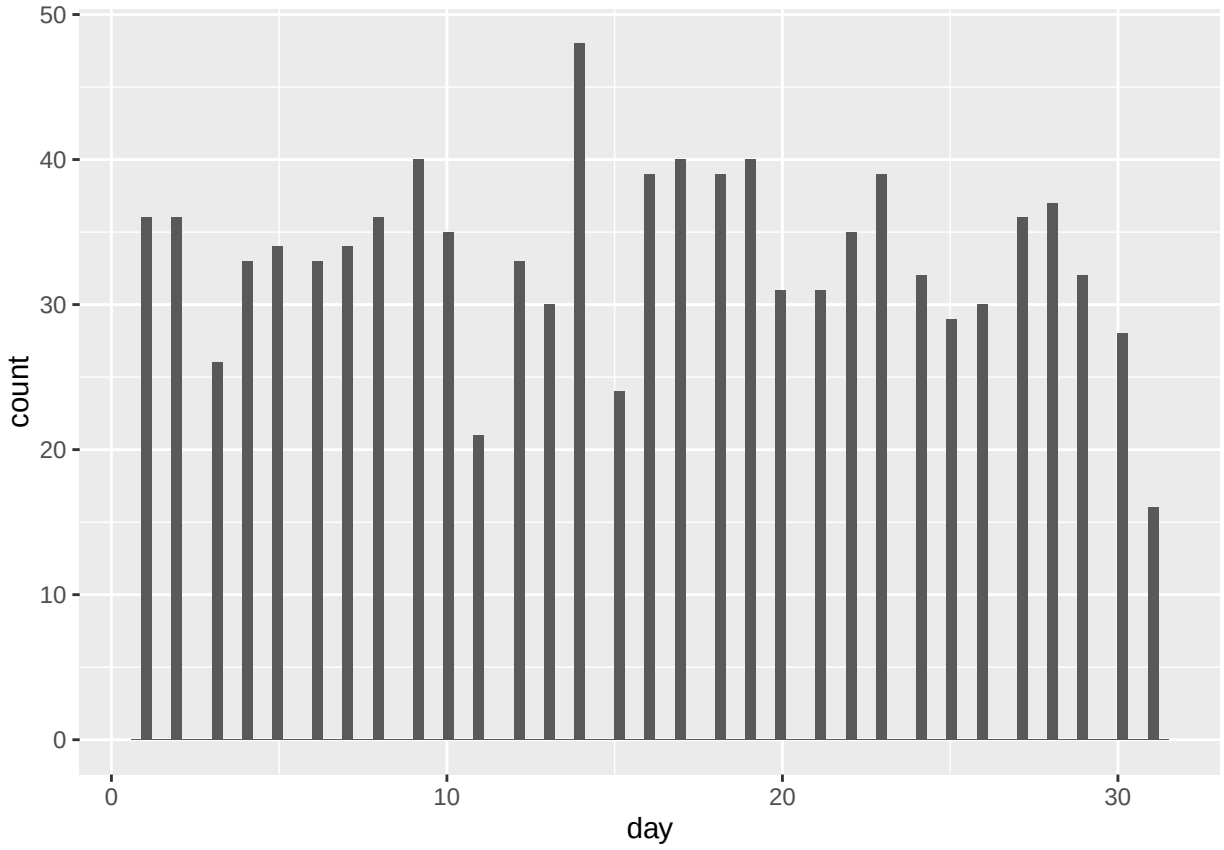
```
qplot(x=month, data=dates, geom="histogram", bins=100)
```



As we can see from the graph, the birthdays are pretty evenly distributed.

Lets Look at the histogram for days:

```
qplot(x=day, data=dates, geom="histogram", bins=100)
```

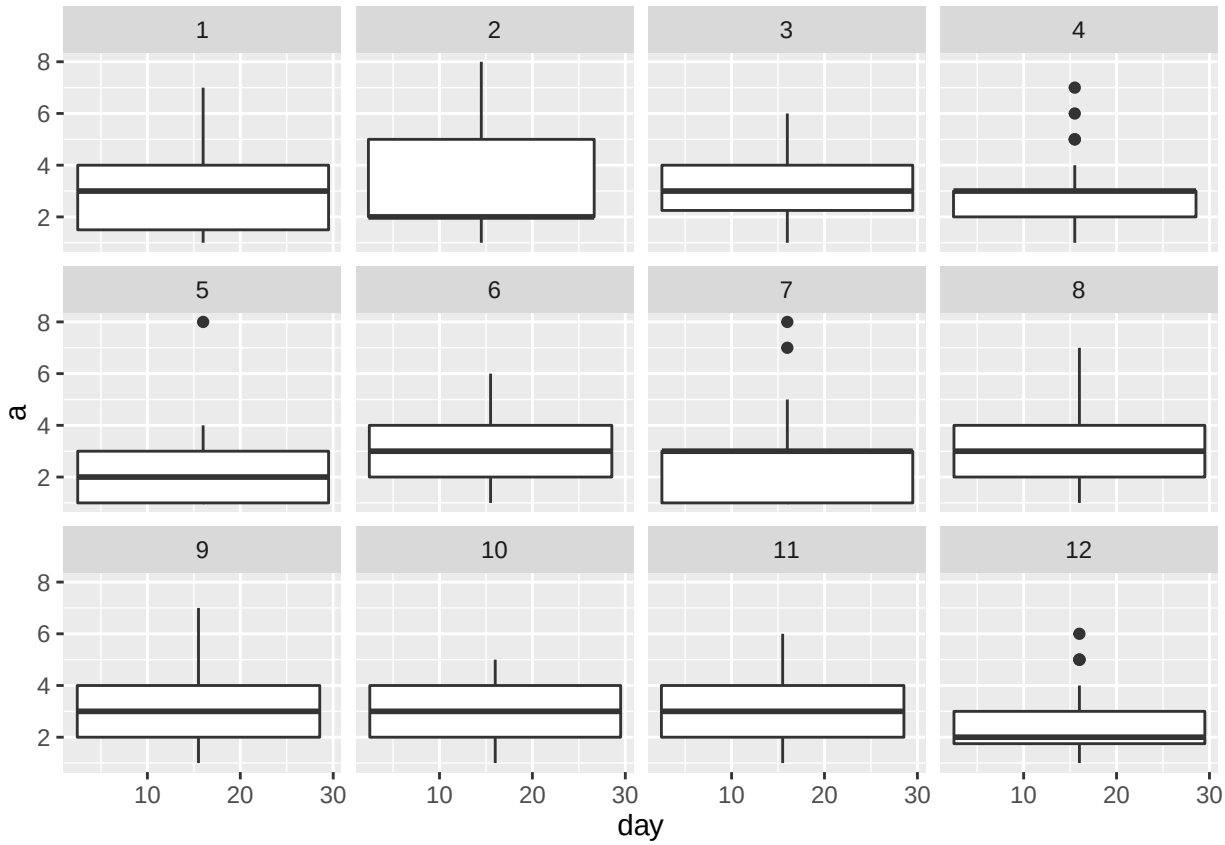


Observing the hostogram for days, looks like there are few dips and a peak. Peak on 14th and dip on 31st and 11th. Dip on 31st is expected as it doesn't appear in all the months.

create an aggregated dataset for total number of birthdays for each day of the month using dplyr

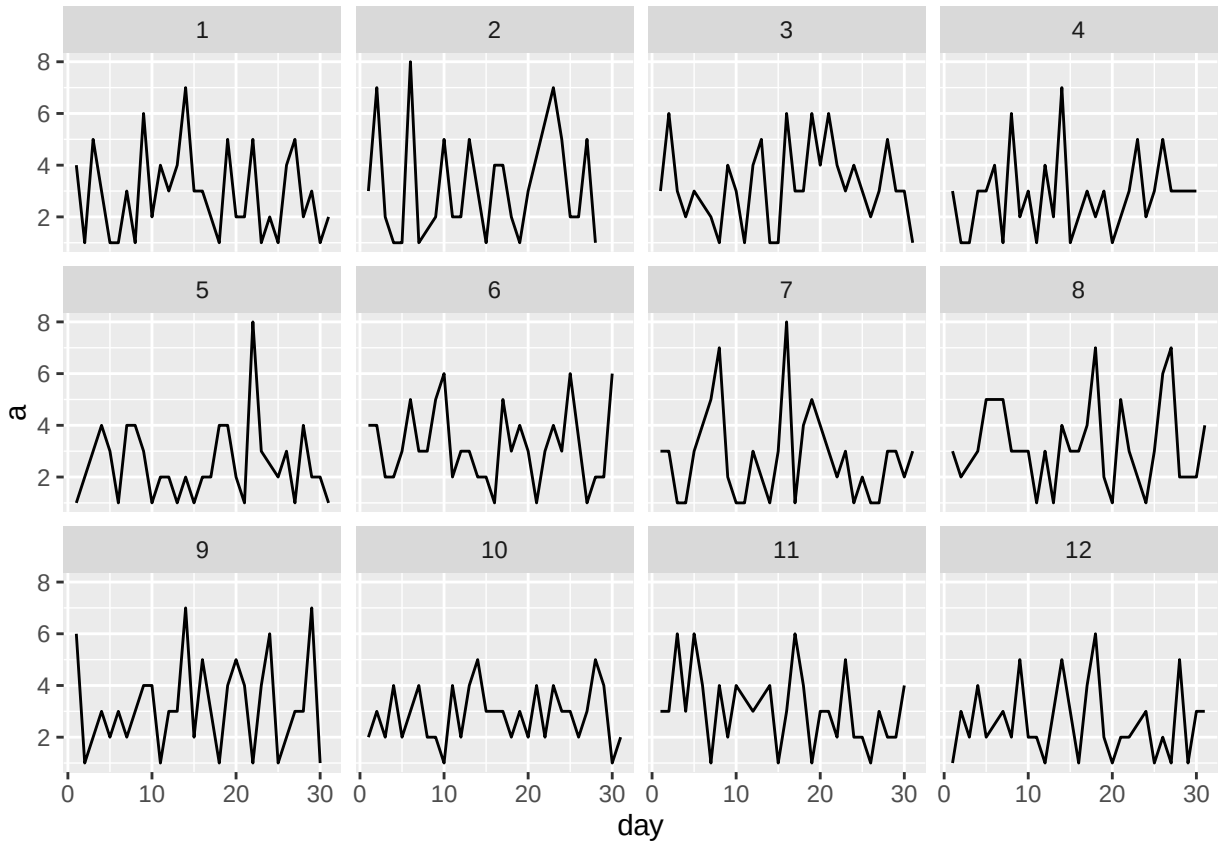
```
dates_aggr <- dates %>% group_by(month, day) %>% summarize(a=n())


ggplot(dates_aggr, aes(day,a)) + geom_boxplot() + facet_wrap(~month)
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



let us see which month has a highest number of birthdays on a particular day:

```
dates_aggr %>% ggplot() + geom_line(aes(x=day, y=a)) + facet_wrap(~month)
```



The graphs indicate that few months have days that have as many as 8 birthdays on the same day. February, May and July.