DATA MINING AND BUSINESS ANALYTICS WITH R

FdR

9 March 2015

This document intends to follow the book **DATA MINING AND BUSINESS ANALYTICS WITH R** from *Johannes Ledolter*. We are taking his work and adapting it to fit the dplyr + ggplot2 + tidyr set of libraries, as well as others when appropriate.

Chapter 2. Processing the Information and Getting to Know Your Data

2.1 2006 Birth data.

We first load the library *nutshell* which contains our dataset, then load it and have a quick look at it.

```
library(nutshell)
```

```
## Loading required package: nutshell.bbdb
## Loading required package: nutshell.audioscrobbler
```

```
data(births2006.smpl)
str(births2006.smpl)
```

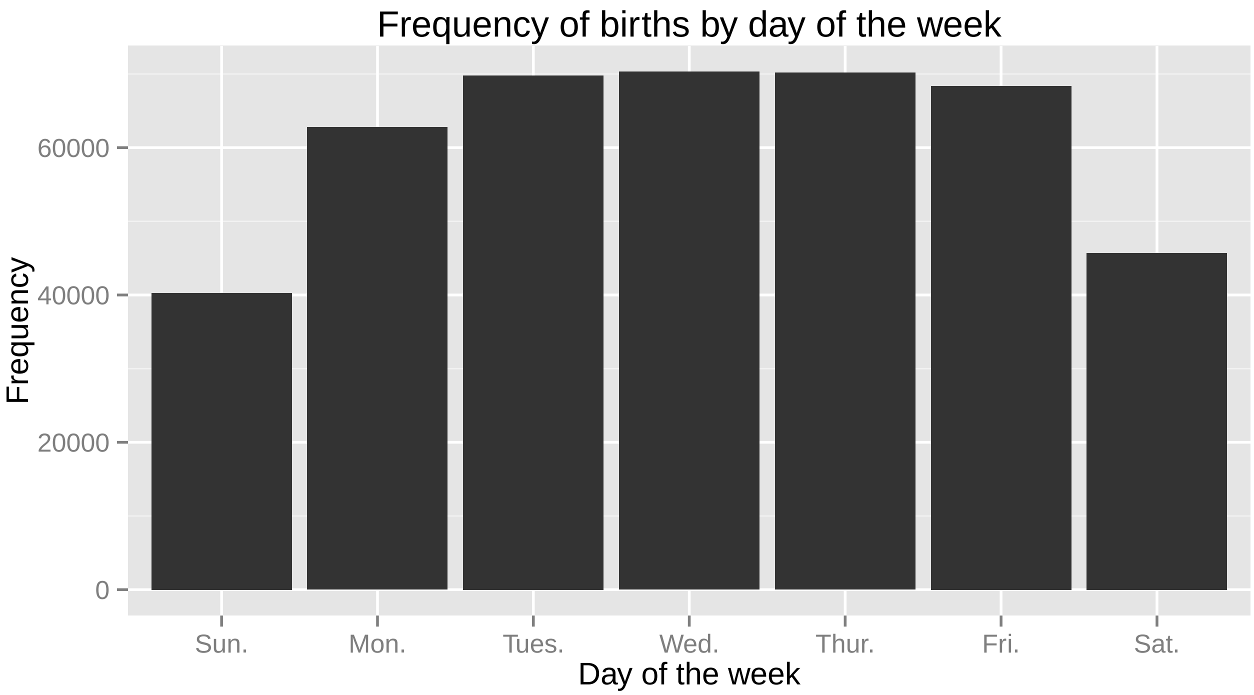
```
## 'data.frame':    427323 obs. of  13 variables:
##  $ DOB_MM   : int  9 2 2 10 7 3 5 4 10 4 ...
##  $ DOB_WK   : int  1 6 2 5 7 3 2 7 3 4 ...
##  $ MAGER    : int  25 28 18 21 25 28 33 31 18 24 ...
##  $ TBO_REC  : int  2 2 2 2 1 3 2 3 1 2 ...
##  $ WTGAIN   : int  NA 26 25 6 36 35 26 25 46 43 ...
##  $ SEX      : Factor w/ 2 levels "F","M": 1 2 1 2 2 2 2 1 1 2 ...
##  $ APGAR5   : int  NA 9 9 9 10 8 9 9 9 9 ...
##  $ DMEDUC   : Factor w/ 18 levels "1 year of college",..: 18 4 18 18 6 18 18 4 18 6 ...
##  $ UPREVIS  : int  10 10 14 22 15 18 10 19 15 13 ...
##  $ ESTGEST  : int  99 37 38 38 40 39 38 38 40 40 ...
##  $ DMETH_REC: Factor w/ 3 levels "C-section","Unknown",..: 3 3 3 3 3 3 1 1 1 3 ...
##  $ DPLURAL  : Factor w/ 5 levels "1 Single","2 Twin",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ DBWT     : int  3800 3625 3650 3045 3827 3090 3430 3204 3227 3459 ...
```

```
head(births2006.smpl)
```

```
##         DOB_MM DOB_WK MAGER TBO_REC WTGAIN SEX APGAR5
## 591430       9      1    25       2     NA   F     NA
## 1827276      2      6    28       2     26   M      9
## 1705673      2      2    18       2     25   F      9
## 3368269     10      5    21       2      6   M      9
## 2990253      7      7    25       1     36   M     10
## 966967       3      3    28       3     35   M      8
##                          DMEDUC UPREVIS ESTGEST DMETH_REC  DPLURAL DBWT
## 591430                     NULL      10      99   Vaginal 1 Single 3800
## 1827276     2 years of college      10      37   Vaginal 1 Single 3625
## 1705673                    NULL      14      38   Vaginal 1 Single 3650
## 3368269                    NULL      22      38   Vaginal 1 Single 3045
## 2990253 2 years of high school      15      40   Vaginal 1 Single 3827
## 966967                     NULL      18      39   Vaginal 1 Single 3090
```

Our first graph is just about the frequency of birth in function of the day of the week.

```
library(ggplot2)
ggplot(births2006.smpl, aes(x = DOB_WK)) +
  geom_bar() +
  scale_x_discrete(labels = c("Sun.", "Mon.", "Tues.", "Wed.", "Thur.", "Fri.", "Sat."),
                   limits = c(1:7)) +
  labs(title = "Frequency of births by day of the week",
       x = "Day of the week", y = "Frequency")
```



They are clearly less birth on the weekend!

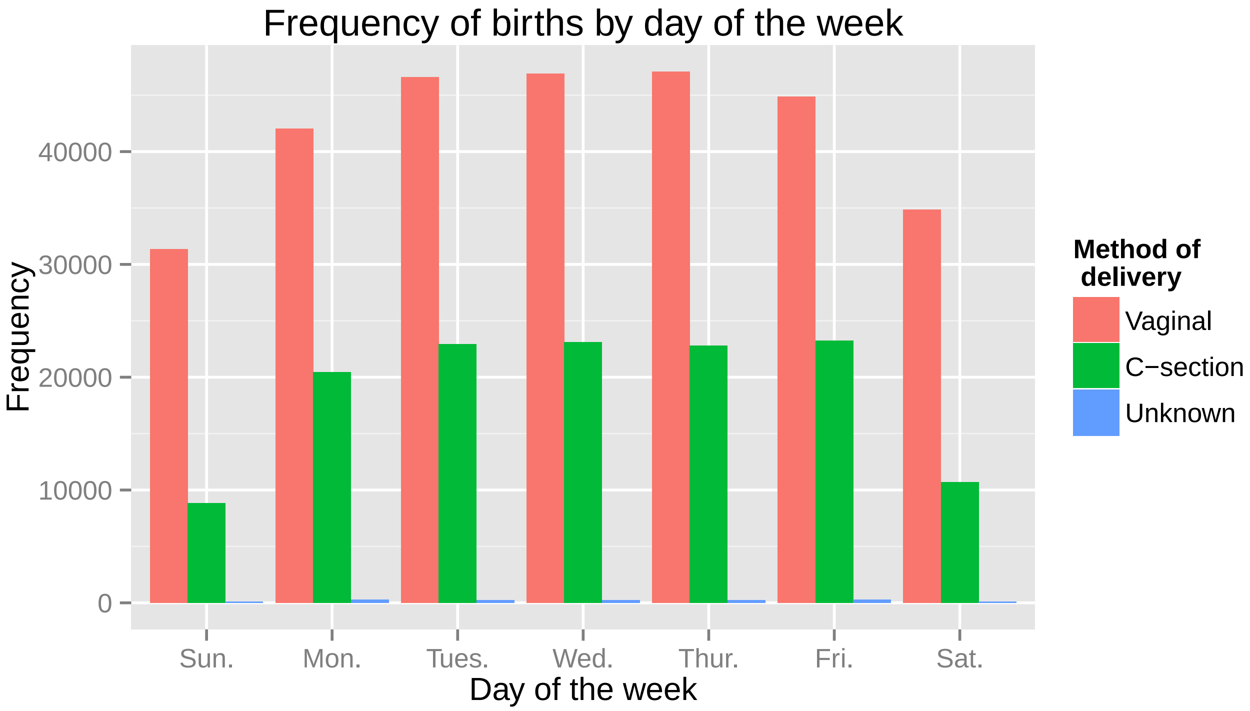
Or we can segregate by method of delivery and graph it that way.

```
table(births2006.smpl$DOB_WK, births2006.smpl$DMETH_REC)
```

```
##
##     C-section Unknown Vaginal
##   1      8836      90   31348
##   2     20454     272   42031
##   3     22921     247   46607
##   4     23103     252   46935
##   5     22825     258   47081
##   6     23233     289   44858
##   7     10696     109   34878
```

We first re-order the levels in the `DMETH_REC` variable, so that the plot look pretty normal.

```
births2006.smpl$DMETH_REC <- factor(births2006.smpl$DMETH_REC,
                                    levels = c("Vaginal", "C-section", "Unknown"))
ggplot(births2006.smpl, aes(x = DOB_WK, fill = DMETH_REC)) +
  geom_bar(position = "dodge") +
  scale_x_discrete(labels = c("Sun.", "Mon.", "Tues.", "Wed.", "Thur.", "Fri.", "Sat."),
                   limits = c(1:7)) +
  labs(title = "Frequency of births by day of the week",
       x="Day of the week", y = "Frequency", fill = "Method of \n delivery")
```

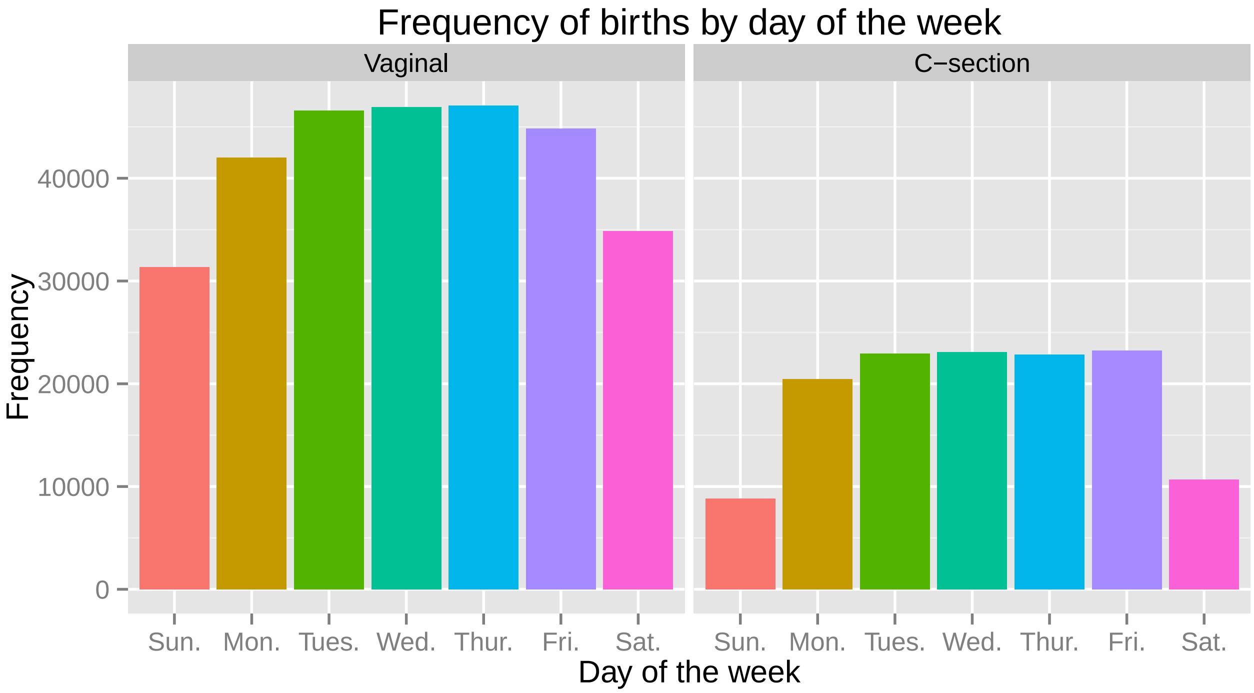


On the next graph, we are using `facet_grid()` to make 2 graphs. There is a bit of plumbing to do first tough.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data <- as.data.frame(table(births2006.smpl$DOB_WK, births2006.smpl$DMETH_REC))
data %>%
  filter(Var2 != "Unknown") %>%
  ggplot(aes(x=Var1, y=Freq, fill = Var1)) +
  geom_bar(stat="identity") + facet_grid(.~ Var2) +
  guides(fill=F) +
  scale_x_discrete(labels = c("Sun.", "Mon.", "Tues.", "Wed.", "Thur.", "Fri.", "Sat.")) +
  labs(title = "Frequency of births by day of the week",
       x="Day of the week", y = "Frequency")
```



I think it might be slightly more interesting to actually compare the percentage of birth by vaginal or C-section on each day of the week.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.