|  |
| --- |
| itle: “R Notebook” |
| utput: html\_notebook |

This data set consists of 427,323 records and 13 variables, including the day of birth according to the month and the day of week (DOB\_MM, DOB\_WK) and other variables like the estimated gestation age in weeks (ESTGEST). We will analyze this data in order to to unearth some information / get to know the data by doing some visualizations.

Let’s glimpse at the first 5 rows of our data to get a sens of our data

library(lattice)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.5 v dplyr 1.0.3  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(nutshell)

## Loading required package: nutshell.bbdb

## Loading required package: nutshell.audioscrobbler

data(births2006.smpl)  
birth2006<- births2006.smpl  
dim(birth2006)

## [1] 427323 13

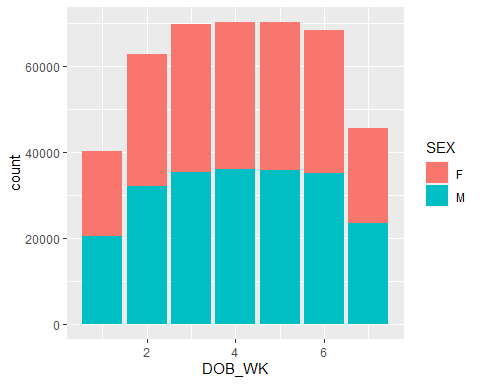
birth2006[1:5,]

## DOB\_MM DOB\_WK MAGER TBO\_REC WTGAIN SEX APGAR5 DMEDUC  
## 591430 9 1 25 2 NA F NA NULL  
## 1827276 2 6 28 2 26 M 9 2 years of college  
## 1705673 2 2 18 2 25 F 9 NULL  
## 3368269 10 5 21 2 6 M 9 NULL  
## 2990253 7 7 25 1 36 M 10 2 years of high school  
## UPREVIS ESTGEST DMETH\_REC DPLURAL DBWT  
## 591430 10 99 Vaginal 1 Single 3800  
## 1827276 10 37 Vaginal 1 Single 3625  
## 1705673 14 38 Vaginal 1 Single 3650  
## 3368269 22 38 Vaginal 1 Single 3045  
## 2990253 15 40 Vaginal 1 Single 3827

## We will look at the frequency of births per day of the week .

This below graphs will help us understand the distribution of our data; We will take a look at the distribution of date of birth in order to know in which day of the week there were more birth: We see that most the women gave birth between 3rd and 5th days of the week. For instance in the 3rd (Wednesday), 4th (Thursday) and 5th (Friday) , there were almost 7000 birth

library(ggplot2)  
  
ggplot(birth2006, aes(DOB\_WK ,fill =SEX)) +  
 geom\_bar()

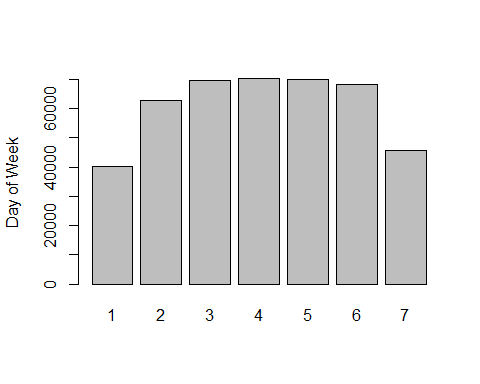


As we can see Thursday has the highest number of birth with a record of *70290* while Monday has the lowest birth with a record of *40274*

birth\_dow<-table(birth2006$DOB\_WK)  
birth\_dow

##   
## 1 2 3 4 5 6 7   
## 40274 62757 69775 70290 70164 68380 45683

barplot(birth\_dow , ylab="Day of Week")



## Birth Delivery method:

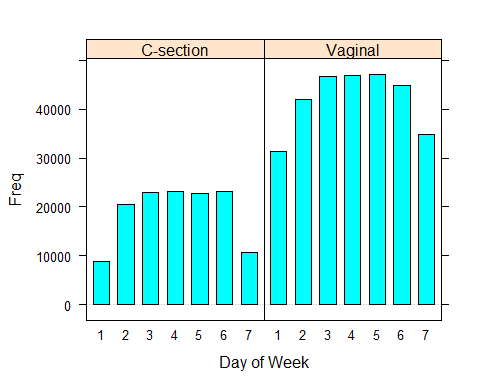
We will classify the ways of birth delivery according to the the week and the method of the delivery

dob\_tbl<-table(WK=birth2006$DOB\_WK, MM=birth2006$DMETH\_REC)  
dob\_tbl

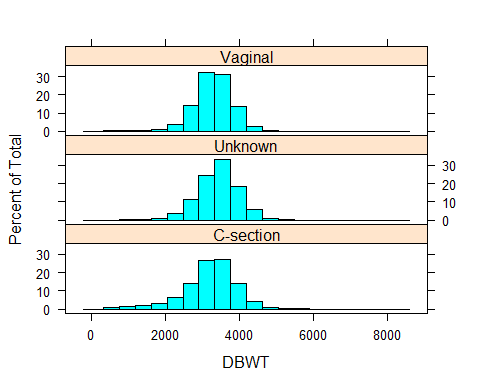
## MM  
## WK C-section Unknown Vaginal  
## 1 8836 90 31348  
## 2 20454 272 42031  
## 3 22921 247 46607  
## 4 23103 252 46935  
## 5 22825 258 47081  
## 6 23233 289 44858  
## 7 10696 109 34878

As we can see there, vaginal delivery method was the method that was mostly used in 2006.

dob\_tbl\_f <- as.data.frame.matrix(dob\_tbl, stringsAsFactors = default.stringsAsFactors())  
  
barchart(dob\_tbl[,-2],horizontal=FALSE,groups=FALSE,xlab="Day of Week")

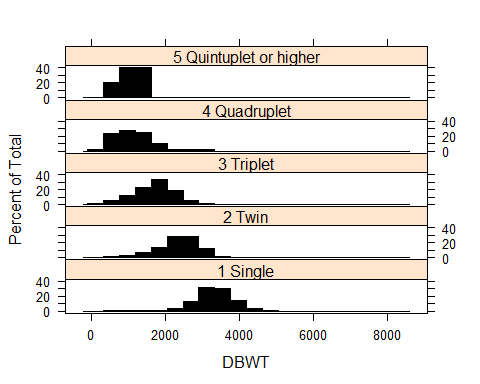


histogram(~DBWT|DMETH\_REC, data = birth2006, layout=c(1,3))

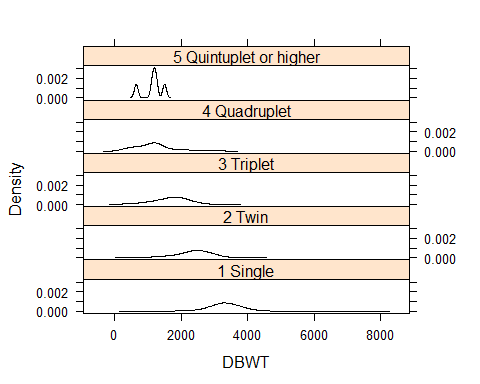


The above bar chart of the frequencies of births according to the day of week of the birth shows that fewer births take place during the weekend (days 1=Sunday, 2=Monday, …,7=Saturday of DOB\_WK). This may have to do with the fact that many babies are delivered by cesarean section, and those deliveries are usually scheduled during the week and not on weekends

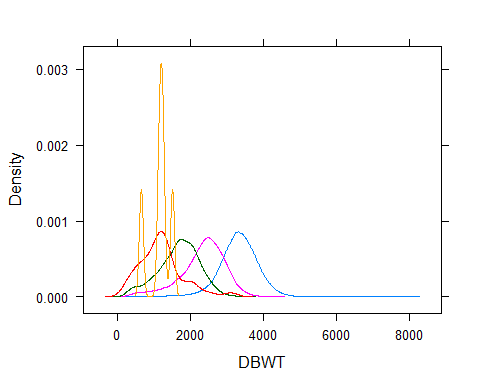
histogram(~DBWT|DPLURAL, data=birth2006, layout=c(1,5), plot.points=FALSE,col="black")



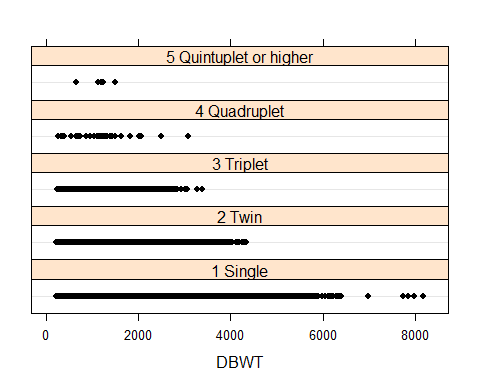
densityplot(~DBWT|DPLURAL, data=birth2006,layout=c(1,5), plot.points=FALSE, col="black")



densityplot(~DBWT, groups = DPLURAL, data=birth2006, plot.points=FALSE, size=3)

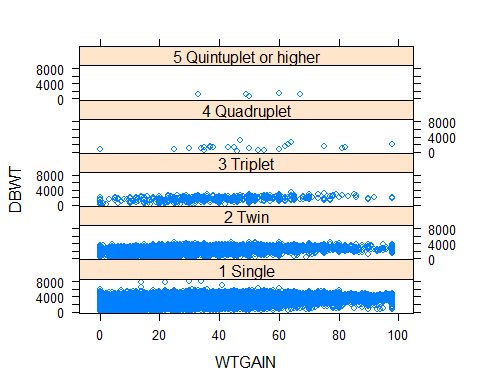


dotplot(~DBWT|DPLURAL, data=birth2006, layout=c(1,5), plot.points=FALSE,col="black")



This Scatter plots is shown for birth weight against weight gain: We see that single & Twin birth babies’ weight the most. In addition, the distribution the single birth is normal, Twin is almost normal too; but triplet, quadruplet, and quintuplet distributions are left skewed, probably those babies don’t weight as the single & Twin babies. We can conclude that twin, triplet, quadruplet, and quintuplet birht have lower birth rates

#xyplot(DBWT~WTGAIN, data=birth2006)  
xyplot(DBWT~WTGAIN|DPLURAL, data = birth2006, layout=c(1,5))



#xyplot(DBWT~WTGAIN|DPLURAL, data=birth2006, layout=c(1,5))

In this section we will Calculate the avarage birth weight for males & females

fac<-factor(birth2006$DPLURAL)  
res<-birth2006$DBWT  
  
avg\_Per\_Gender<- tapply(birth2006$DBWT, INDEX = list(birth2006$DPLURAL, birth2006$SEX), FUN = mean, na.rm=T)  
avg\_Per\_Gender

## F M  
## 1 Single 3242.302 3351.637  
## 2 Twin 2279.508 2373.819  
## 3 Triplet 1697.822 1655.348  
## 4 Quadruplet 1319.556 1085.000  
## 5 Quintuplet or higher 1007.667 1345.500

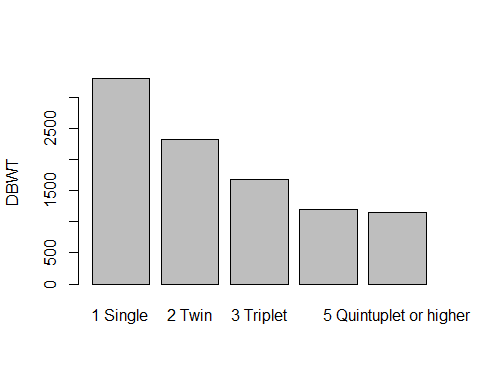
Here we will finally Calculate the avarage birth weight for males & females

This below graph illustrate the conclusion we made earlier. Single babies have almost twice the weight of Triplet.

fac<-factor(birth2006$DPLURAL)  
res<-birth2006$DBWT  
avg\_Per\_Birth<-tapply(res, fac, mean, na.rm=T)  
avg\_Per\_Birth

## 1 Single 2 Twin 3 Triplet   
## 3298.263 2327.478 1677.017   
## 4 Quadruplet 5 Quintuplet or higher   
## 1196.105 1142.800

barplot(avg\_Per\_Birth, ylab="DBWT")



Frequency distribution table of estimated gestation period indicates that “99” is the code for “UNKNOW”. The following calculations will omit all records with unknown gestation period(vaue 99)

avg\_Per\_Gender<-table(birth2006$ESTGEST)  
new<-birth2006[birth2006$ESTGEST != 99,]  
GestTable<- table(new$ESTGEST)  
GestTable

##   
## 12 15 17 18 19 20 21 22 23 24 25   
## 1 2 18 43 69 116 162 209 288 401 445   
## 26 27 28 29 30 31 32 33 34 35 36   
## 461 566 670 703 1000 1243 1975 2652 4840 7954 15874   
## 37 38 39 40 41 42 43 44 45 46 47   
## 33310 76794 109046 84890 23794 1931 133 32 6 5 5   
## 48 51   
## 2 1