Introduction to RSQLite Database

Michelle Norris

November 11, 2017

Introduction

The website Kaggle.com has a collection of interesting datasets. One is the World Development Indicators dataset at this url link. Here is the description:

> "The World Development Indicators from the World Bank contain over a thousand annual indicators of economic development from hundreds of countries around the world. Here's a list of the available indicators along with a list of the available countries."

For example, this data includes the life expectancy at birth from many countries around the world.

Let's go to the url and check out a few cool graphs using this data.

When you uncompress the data, you obtain many files. The primary data are stored in a 1.4 GB file called **database.sqlite**. There are auxiliary files which list information like: all countries in the database, all economic indicators in the database, notes about the data, etc.

Small Database Example

Let's work with a small database first to get a feel for RSQLite.

```
library(RSQLite)
```

```
## Warning: package 'RSQLite' was built under R version 3.4.1
```

```
myconnection = dbConnect(drv=SQLite(),":memory:") # sets up a temporary in-memory database

str(myconnection)
```

```
## Formal class 'SQLiteConnection' [package "RSQLite"] with 6 slots
##   ..@ ptr                :<externalptr>
##   ..@ dbname             : chr ":memory:"
##   ..@ loadable.extensions: logi TRUE
##   ..@ flags              : int 70
##   ..@ vfs                : chr ""
##   ..@ ref                :<environment: 0x00000000092cd3e0>
```

```
dbWriteTable(con=myconnection,name="mtcarsdata",value=mtcars)  # don't save this to a named object; mtc

dbListTables(myconnection)  #see all tables in myconnection
```

```
## [1] "mtcarsdata"
```

```
dbListFields(myconnection,"mtcarsdata")  #list fields=column names
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Basic SQL Query

A query is used to extract portions of a database or to summarize records in a database. The syntax of the query to extract specific columns is: **SELECT column-list , FROM table-name** . This is SQL (Structured Query Language) syntax and needs to be embedded in an R command for query.

```
head(dbGetQuery(myconnection,"SELECT disp,cyl FROM mtcarsdata"))
```

```
##   disp cyl
## 1  160   6
## 2  160   6
## 3  108   4
## 4  258   6
## 5  360   8
## 6  225   6
```

```
head(dbGetQuery(myconnection,"SELECT * FROM mtcarsdata"))  #use the wildcard * to get all columns
```

```
##    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

RSQLite with a Large database

Let's use RSQLite to read in the database, get information from the data, and run some queries. First, put the database in your working directory. Next, we must create a connection to the database.

```
dbDisconnect(myconnection)  # disconnect the mtcars database
myconnection2 = dbConnect(drv=SQLite(), dbname="C:\\Users\\sac87931\\Documents\\kaggle\\world-developme
str(myconnection2)
```

```
## Formal class 'SQLiteConnection' [package "RSQLite"] with 6 slots
##   ..@ ptr                :<externalptr>
##   ..@ dbname             : chr "C:\\Users\\sac87931\\Documents\\kaggle\\world-development-indicators
##   ..@ loadable.extensions: logi TRUE
##   ..@ flags              : int 70
##   ..@ vfs                : chr ""
##   ..@ ref                :<environment: 0x000000000bc235a0>
```

```
dbListTables(myconnection2)
```

```
## [1] "Country"      "CountryNotes" "Footnotes"    "Indicators"
## [5] "Series"       "SeriesNotes"
```

Each of the tables listed is a data frame. Let's look at the fields (or column names in a few tables).

```
head(dbListFields(myconnection2,"Indicators"))
```

```
## [1] "CountryName"   "CountryCode"   "IndicatorName" "IndicatorCode"
## [5] "Year"          "Value"
```

```
head(dbListFields(myconnection2,"Country"))
```

```
## [1] "CountryCode"  "ShortName"    "TableName"    "LongName"
```

```
## [5] "Alpha2Code"   "CurrencyUnit"
```

Extract Life Expectancy data by gender and Create a Graph

Now we extract data on life expectancy for females and males in the US and construct a time series graph.

```
#female life expectancy at birth
results.female.life <- dbSendQuery(myconnection2, "SELECT * FROM Indicators WHERE CountryCode = 'USA' AN

tempp = dbFetch(results.female.life, n=Inf)
head(tempp)
```

```
##     CountryName CountryCode                             IndicatorName
## 1 United States         USA Life expectancy at birth, female (years)
## 2 United States         USA Life expectancy at birth, female (years)
## 3 United States         USA Life expectancy at birth, female (years)
## 4 United States         USA Life expectancy at birth, female (years)
## 5 United States         USA Life expectancy at birth, female (years)
## 6 United States         USA Life expectancy at birth, female (years)
##       IndicatorCode Year Value
## 1 SP.DYN.LE00.FE.IN 1960  73.1
## 2 SP.DYN.LE00.FE.IN 1961  73.6
## 3 SP.DYN.LE00.FE.IN 1962  73.5
## 4 SP.DYN.LE00.FE.IN 1963  73.4
## 5 SP.DYN.LE00.FE.IN 1964  73.7
## 6 SP.DYN.LE00.FE.IN 1965  73.8
```

```
class(tempp)  # result is a dataframe
```

```
## [1] "data.frame"
```

```
dbClearResult(results.female.life) # releases memory

plot(tempp$Year,tempp$Value,type="l",ylim=c(60,80),ylab="Life Expectancy",xlab='Year')


#male life expectancy at birth
results.male.life <- dbSendQuery(myconnection2, "SELECT * FROM Indicators WHERE CountryCode = 'USA' AND
tempp = dbFetch(results.male.life, n=Inf)
head(tempp)
```

```
##     CountryName CountryCode                           IndicatorName
## 1 United States         USA Life expectancy at birth, male (years)
## 2 United States         USA Life expectancy at birth, male (years)
## 3 United States         USA Life expectancy at birth, male (years)
## 4 United States         USA Life expectancy at birth, male (years)
## 5 United States         USA Life expectancy at birth, male (years)
## 6 United States         USA Life expectancy at birth, male (years)
##       IndicatorCode Year Value
## 1 SP.DYN.LE00.MA.IN 1960  66.6
## 2 SP.DYN.LE00.MA.IN 1961  67.1
## 3 SP.DYN.LE00.MA.IN 1962  66.9
## 4 SP.DYN.LE00.MA.IN 1963  66.6
## 5 SP.DYN.LE00.MA.IN 1964  66.8
## 6 SP.DYN.LE00.MA.IN 1965  66.8
```

```
dbClearResult(results.male.life)
points(tempp$Year,tempp$Value,type="l",col="blue",lty=2)
```