Project Notebook

Keith Engwall

1/23/2018

Project Objectives

Create a predictive model for identifying patients with Diabetes

See Capstone Proposal for details.

Project Dataset

Practice Fusion De-Identified Data Set containing EHR data for approximately 10,000 de-identified patients, including data points for diagnoses, medication, transcript data, and lab observations. See the Data Dictionary for details.

Project Notes

Loading data into R

The dataset is contained within an SQLite database file (420.7MB). To load the data into R requires installation and loading of **RSQLite** and **DBI** R packages. One of the dependencies is the tibble package. During the install, I was asked whether to install the binary version (1.3.4) or the source version (1.4.1), which would need compilation. I wasn't comfortable enough to explore compiling R package code yet, so I went with the 1.3.4 version.

I found this brief example of how to connect to and query an SQLite database file to be very helpful in getting up and going quickly.

```
con <- dbConnect(SQLite(), dbname="<SQLite File>.db")
myQuery <- dbSendQuery(con, "<SQL Query>")
```

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

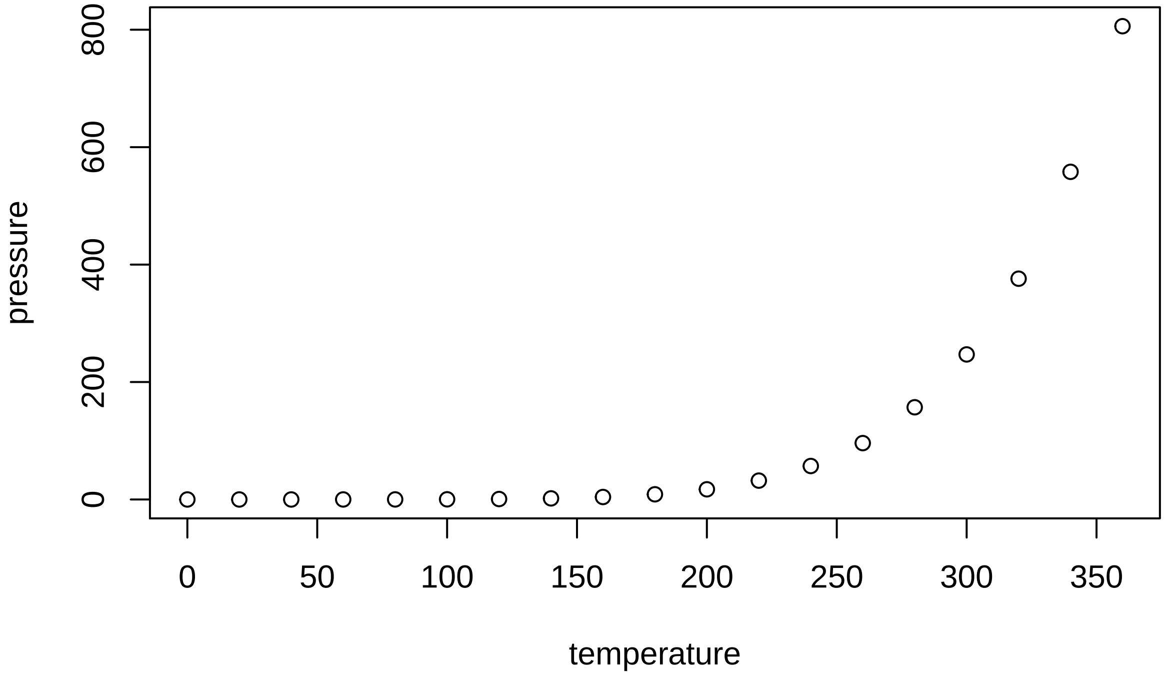
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



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