Data Wrangling Part 1

Tom Skawski II

July 14, 2016

R Markdown

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.

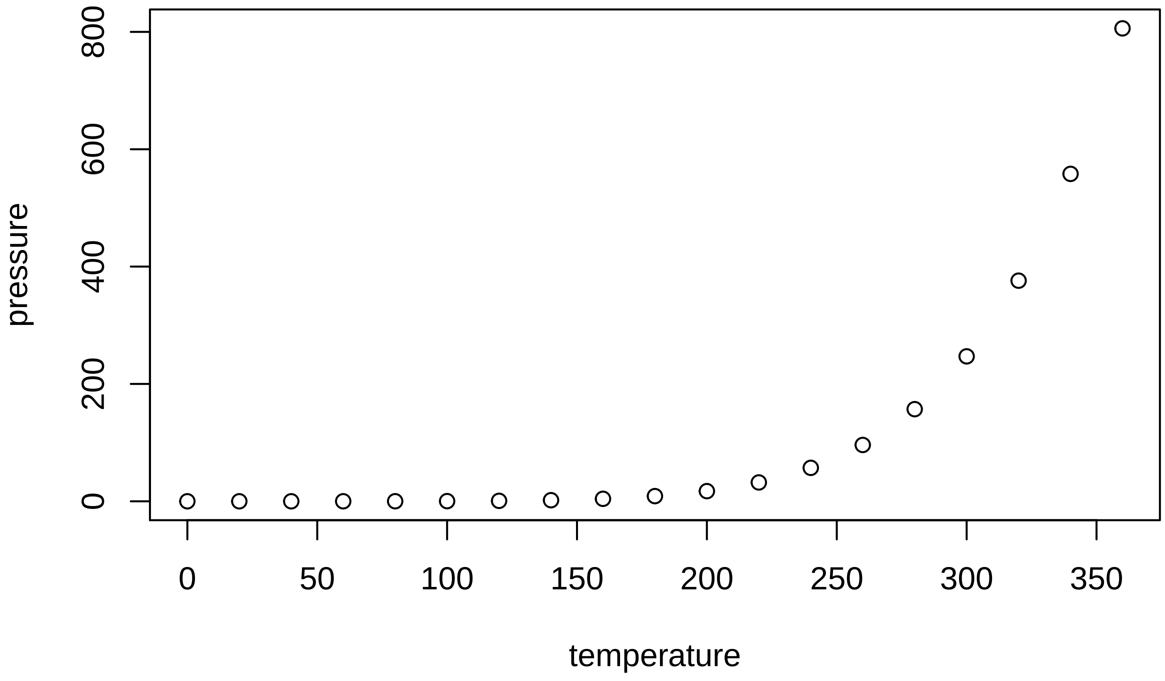
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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.

```
setwd("~tomskawski/Dropbox/Data Science/Data Wrangling")

library("tidyr")
library("dplyr")

## Step 0 - Load data
mydata <- read.csv("refine_original.csv")

# check class of columns. s/b character; if not, convert
str(mydata)
mydata[] <- lapply(mydata, as.character) # [] keeps data.frame, also changes all columns

grep("^p", mydata$company, ignore.case = TRUE, value = TRUE) #value prints data, FALSE gives position
mydata$company <- gsub("\\b(p|f)\\w+", "philips", mydata$company, ignore.case = TRUE) #\\b is word boun
grep("\\ba\\w+", mydata$company, ignore.case = TRUE, value = TRUE)
mydata$company <- gsub("\\ba\\w+", "akzo", mydata$company, ignore.case = TRUE)
mydata$company <- gsub("\\ba\\w+\\s\\w+", "akzo", mydata$company, ignore.case = TRUE)
grep("\\bv\\w+\\s\\w+", mydata$company, ignore.case = TRUE, value = TRUE) #\\s\\w+ is for extra word
mydata$company <- gsub("\\bv\\w+\\s\\w+", "van houten", mydata$company, ignore.case = TRUE) #\\s\\w+ is
grep("\\bu\\w+", mydata$company, ignore.case = TRUE, value = TRUE)
mydata$company <- gsub("\\bu\\w+", "unilever", mydata$company, ignore.case = TRUE)
mydata <- separate(mydata, Product.code...number, c("Prod.code", "Prod.number"), sep = "-")
mydata %>% group_by(company) %>% summarise(country = n())
Prod.code <- c("p", "q", "v", "x")
Prod.cat <- c("Smartphone", "Tablet", "TV", "Laptop")
merge = data_frame(Prod.code, Prod.cat)
mydata <- left_join(mydata, merge)
mydata <- unite(mydata, full_address, address:country, sep = ",") # not sure if this should be a new, c
mydata$company_philips <- as.numeric(mydata$company == "philips")
mydata$company_akzo <- as.numeric(mydata$company == "akzo")
mydata$company_unilever <- as.numeric(mydata$company == "unilever")
mydata$company_van_houten <- as.numeric(mydata$company == "van houten")
mydata$product_smartphone <- as.numeric(mydata$Prod.cat == "Smartphone")
mydata$product_tv <- as.numeric(mydata$Prod.cat == "TV")
mydata$product_laptop <- as.numeric(mydata$Prod.cat == "Laptop")
mydata$product_tablet <- as.numeric(mydata$Prod.cat == "Tablet")
```