demonstration

David LeBauer

February 26, 2016

Intro R

This is demonstration code written while teaching the SWC R Novice Gapminder lesson http://swcarpentry.github.io/r-novice-gapminder/.

```
for ( i in 1:5){
  print(paste("there are ", i, "apples"))
}
```

```
## [1] "there are  1 apples"
## [1] "there are  2 apples"
## [1] "there are  3 apples"
## [1] "there are  4 apples"
## [1] "there are  5 apples"
```

```
#install.packages("ggplot2")
#install.packages("plyr")
#install.packages("dplyr")
#install.packages("gapminder")

mass <- 4
age <- 122
mass2 <- mass * 5

set.seed(1)
matrix(1:50, nrow = 10, ncol = 5)[5,5]
```

```
## [1] 45
```

```
matrix(1:5, nrow = 10, ncol = 10)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1    1    1    1    1    1    1    1    1     1
##  [2,]    2    2    2    2    2    2    2    2    2     2
##  [3,]    3    3    3    3    3    3    3    3    3     3
##  [4,]    4    4    4    4    4    4    4    4    4     4
##  [5,]    5    5    5    5    5    5    5    5    5     5
##  [6,]    1    1    1    1    1    1    1    1    1     1
##  [7,]    2    2    2    2    2    2    2    2    2     2
##  [8,]    3    3    3    3    3    3    3    3    3     3
##  [9,]    4    4    4    4    4    4    4    4    4     4
## [10,]    5    5    5    5    5    5    5    5    5     5
```

```
list(number = 1, list = list(number = 1, letter = "a", truefalsething = TRUE, 1+4i))
```

```
## $number
## [1] 1
##
## $list
## $list$number
## [1] 1
##
## $list$letter
## [1] "a"
##
## $list$truefalsething
## [1] TRUE
##
## $list[[4]]
## [1] 1+4i
```

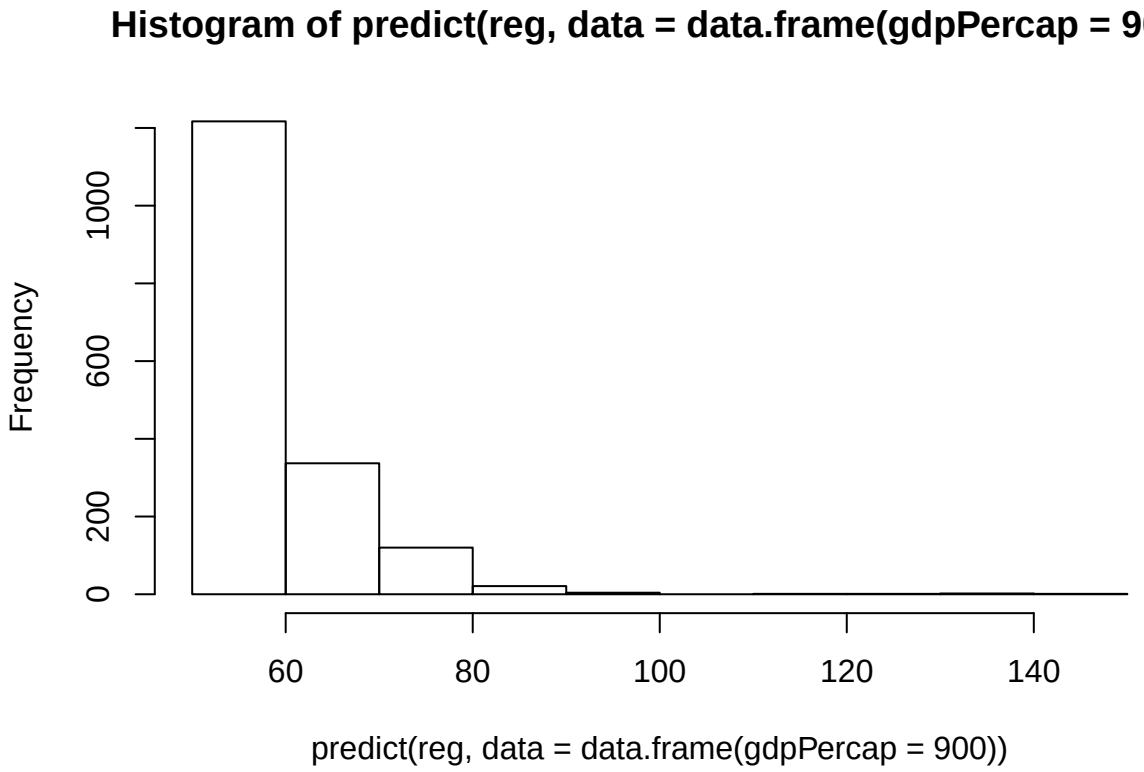
```
mydf <- data.frame(id = c('a', 'b', 'c', 'd', 'e', 'f'),
                   x  = 1:6,
                   y  = 214:219,
                   z  = rnorm(6),
                   e  = LETTERS[6:11])

mydf2 <- cbind(mydf, ans = mydf$x * mydf$y)
mydf3 <- rbind(mydf2, list('z', 2, 222, -2, 'F', 9))
```

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = "z"): invalid factor level, NA
## generated
```

```
gapminder <- read.csv("~/rezbaz/data/gapminder-FiveYearData.csv")

lm(lifeExp ~ continent, data = gapminder)
```

```
##
## Call:
## lm(formula = lifeExp ~ continent, data = gapminder)
##
## Coefficients:
##       (Intercept)  continentAmericas      continentAsia
##             48.87              15.79              11.20
##   continentEurope   continentOceania
##             23.04              25.46
```

```
reg <- lm(lifeExp ~ gdpPercap, data = gapminder)

hist(predict(reg, data = data.frame(gdpPercap = 900)))
```



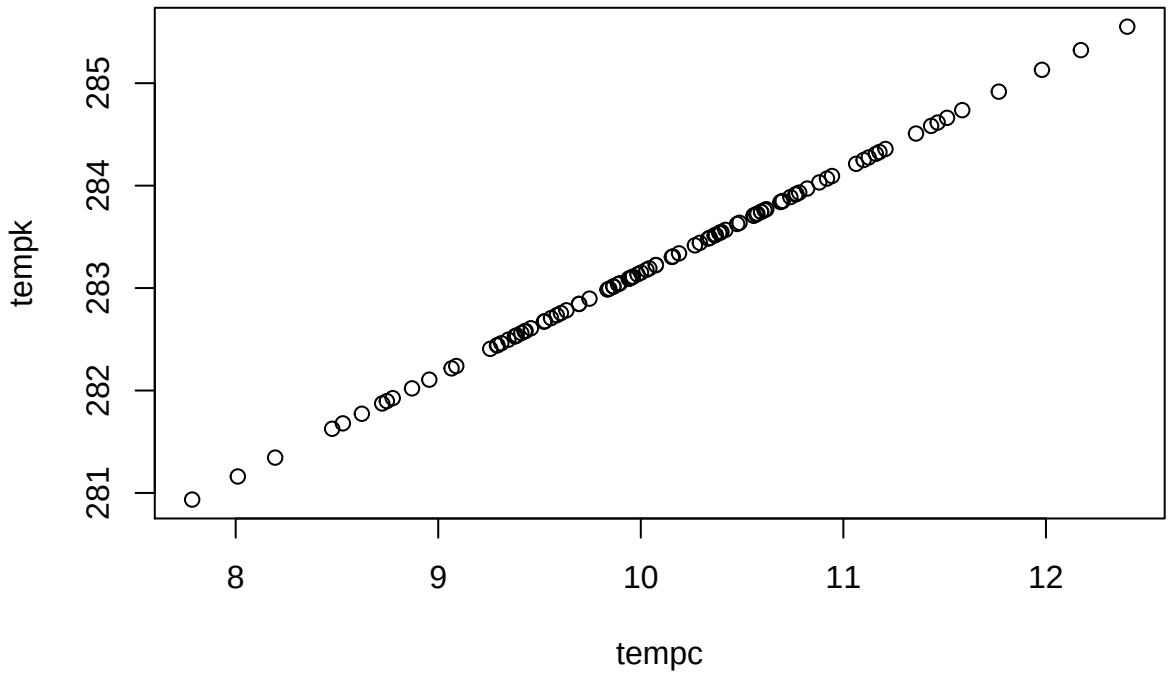

```
range(predict(reg, data = data.frame(gdpPercap = 900)))
```

```
## [1]  54.14002 140.78744
```

Functions

Function 1: Convert C to Kelvin

```
celsius2kelvin <- function(temp_c){
  temp_k <- temp_c + 273.15
  return(temp_k)
}

tempc <- rnorm(100, 10, 1)
tempk <- celsius2kelvin(temp_c = tempc)
plot(tempc, tempk)
```

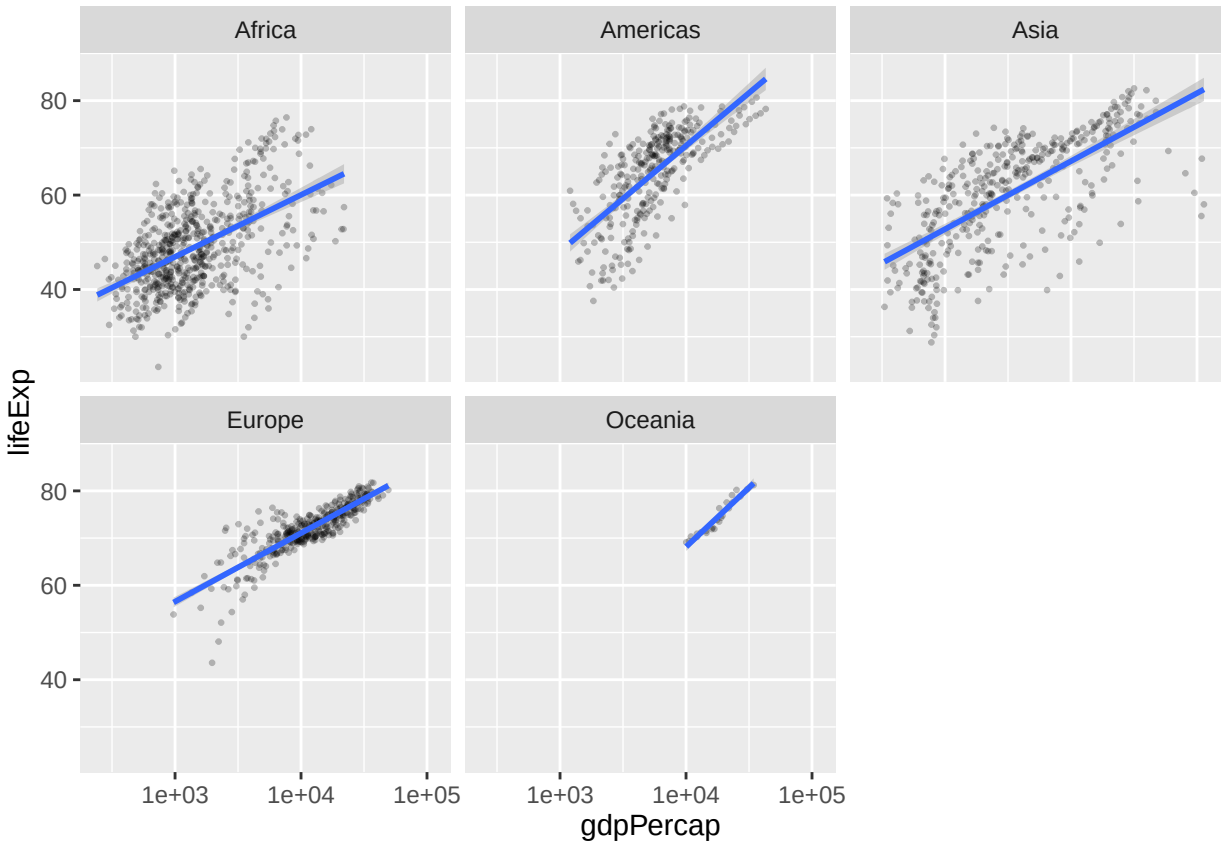


plotting using ggplot

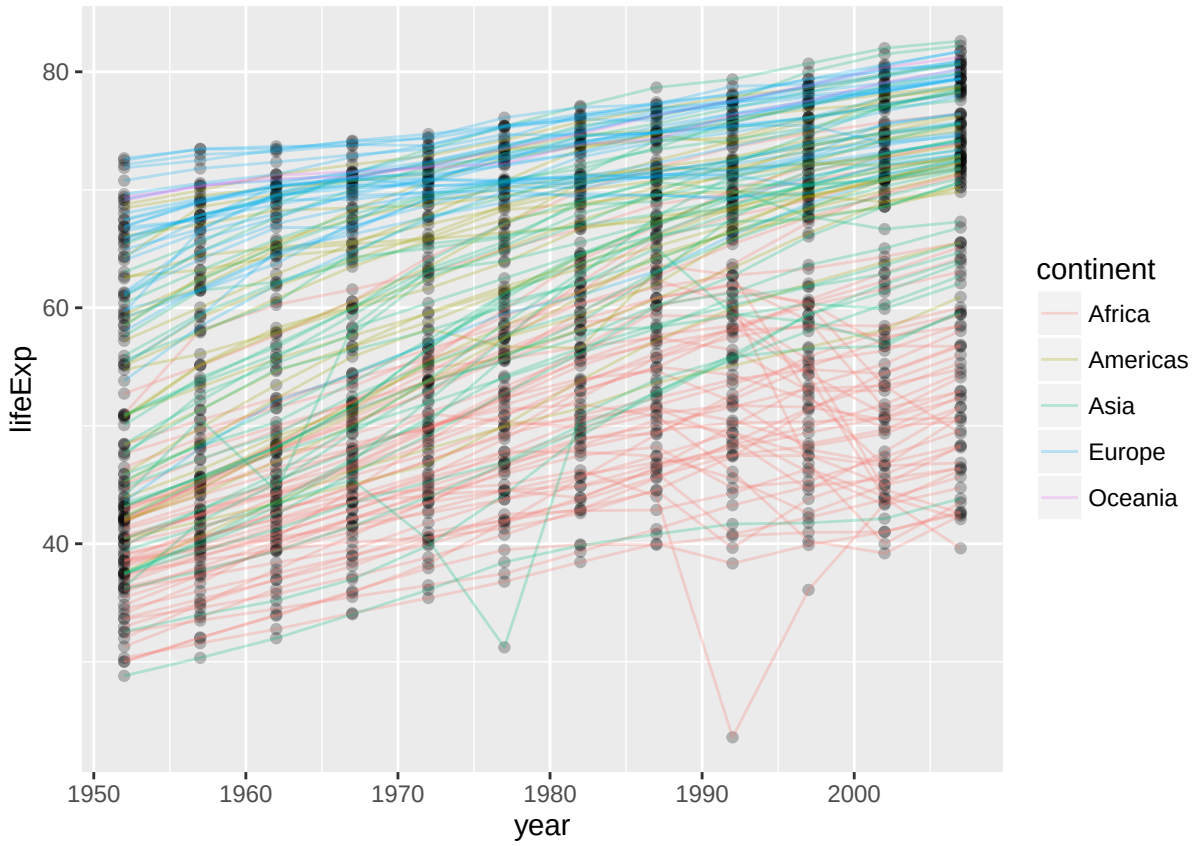
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

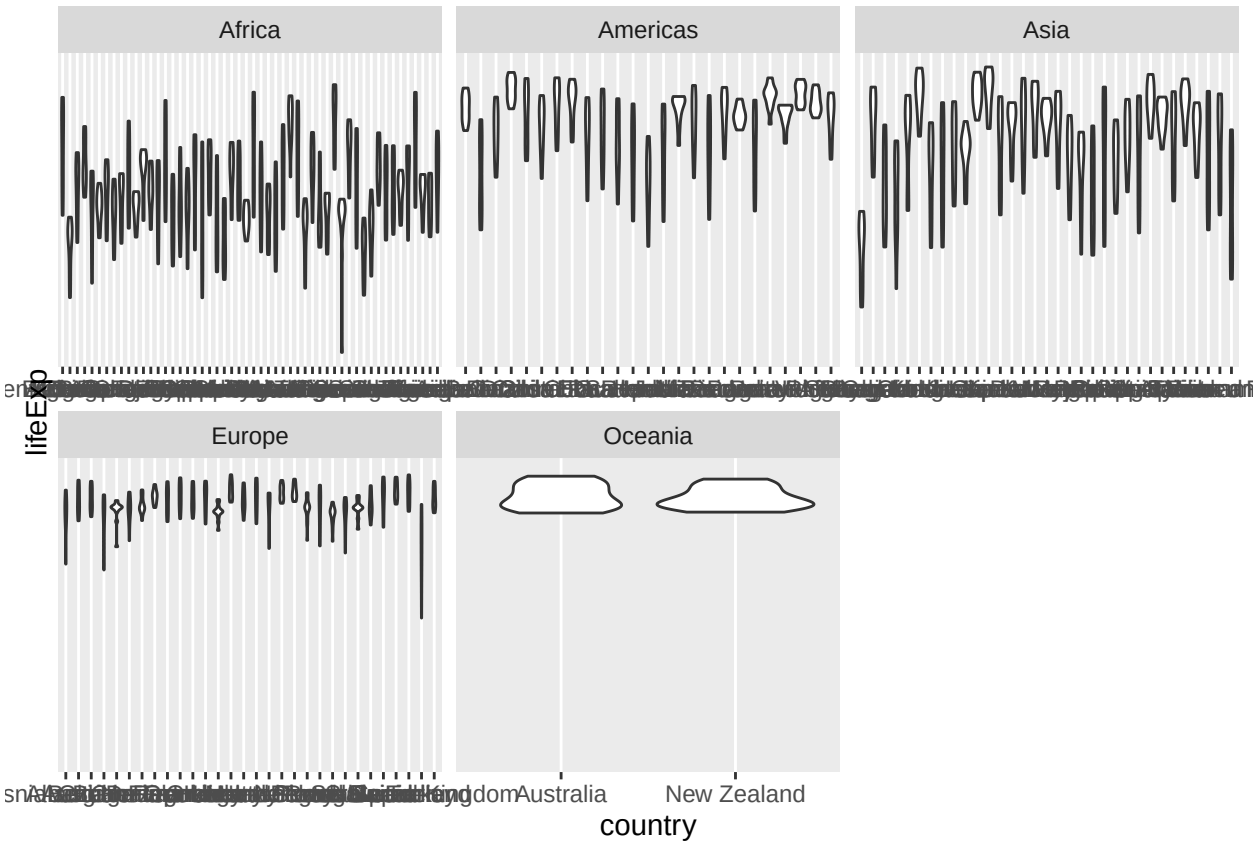
```
gapminder <- read.csv("~/rezbaz/data/gapminder-FiveYearData.csv")

ggplot(data = gapminder, aes(x = gdpPercap, y = lifeExp)) +
  geom_point(alpha = 0.25, size = 0.5) +
  geom_smooth(method = 'lm') +
  facet_wrap(~continent) +
  scale_x_log10()
```



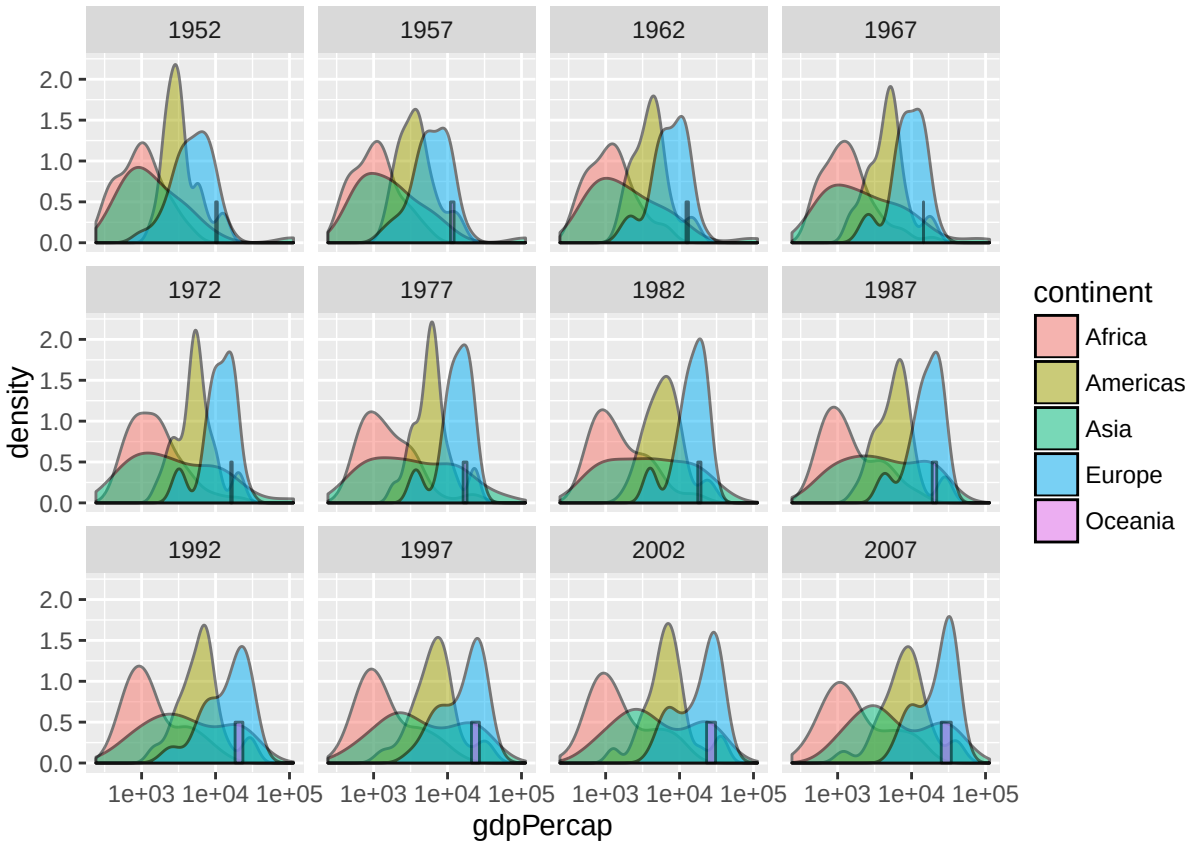

```
ggplot(data = gapminder,
       aes(x = year, y = lifeExp, by = country)) +
  geom_point(alpha = 0.25) +
  geom_line(aes(color = continent), alpha = 0.25)
```




```
ggplot(data = gapminder, aes(x = country, y = lifeExp)) +
  geom_violin() +
  scale_y_log10() +
  facet_wrap(~continent, scales = 'free_x')
```




```
ggplot(data = gapminder, aes(x = gdpPercap, fill = continent)) + geom_density(alpha = 0.5) + facet_wrap
```




```
summary(gapminder)
```

```
##         country          year            pop                continent
##  Afghanistan:  12   Min.   :1952   Min.   :6.001e+04   Africa  :624
##  Albania    :  12   1st Qu.:1966   1st Qu.:2.794e+06   Americas:300
##  Algeria    :  12   Median :1980   Median :7.024e+06   Asia    :396
##  Angola     :  12   Mean   :1980   Mean   :2.960e+07   Europe  :360
##  Argentina  :  12   3rd Qu.:1993   3rd Qu.:1.959e+07   Oceania : 24
##  Australia  :  12   Max.   :2007   Max.   :1.319e+09
##  (Other)    :1632
##     lifeExp        gdpPercap
##  Min.   :23.60   Min.   :   241.2
##  1st Qu.:48.20   1st Qu.:  1202.1
##  Median :60.71   Median :  3531.8
##  Mean   :59.47   Mean   :  7215.3
##  3rd Qu.:70.85   3rd Qu.:  9325.5
##  Max.   :82.60   Max.   :113523.1
##
```

Using databases from R

Here are a few different methods. I prefer the `dplyr` approach because it helps to break down and simplify the syntax of complex operations in SQL. Here we show it with sqlite, but it also works with (almost) any (relational) database manager see `?src_sql`, `?src_mysql`, `src_postgres`.

dplyr

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
surveydb <- src_sqlite("~/swcarpentry/my_project/data/survey.db")
survey <- tbl(surveydb, 'Survey')
class(survey)
```

```
## [1] "tbl_sqlite" "tbl_sql"    "tbl"
```

```
visited <- tbl(surveydb, 'Visited')

### The following three are equivalent

#3 the non piped approach
survey_lakeroe <- filter(survey, person %in% c('lake', 'roe'))

#2 What the above is doing. if survey is not first argument, use '.' where it belongs
survey_lakeroe <- survey %>%
  filter(., person %in% c('lake', 'roe'))

#1 The simple but most common syntax for dplyr, shorthand for above:
survey_lakeroe <- survey %>%
  filter(person %in% c('lake', 'roe'))

## another use of pipes

gapminder %>%
  select(country, gdpPercap) %>%
  filter(country == "Zimbabwe")
```

```
##     country gdpPercap
## 1  Zimbabwe  406.8841
## 2  Zimbabwe  518.7643
## 3  Zimbabwe  527.2722
## 4  Zimbabwe  569.7951
## 5  Zimbabwe  799.3622
## 6  Zimbabwe  685.5877
## 7  Zimbabwe  788.8550
```

```
## 8  Zimbabwe  706.1573
## 9  Zimbabwe  693.4208
## 10 Zimbabwe  792.4500
## 11 Zimbabwe  672.0386
## 12 Zimbabwe  469.7093
```

```
# two equivalent ways of joining

## send an SQL statement:
tbl(surveydb,
    sql("Select * from visited join survey on visited.ident = survey.taken"))
```

```
## Source: sqlite 3.8.6 [~/swcarpentry/my_project/data/survey.db]
## From: <derived table> [?? x 7]
##
##    ident  site      dated taken person quant reading
##    (int) (chr)      (chr) (int)  (chr) (chr)   (dbl)
## 1    619  DR-1 1927-02-08   619   dyer   rad    9.82
## 2    619  DR-1 1927-02-08   619   dyer   sal    0.13
## 3    622  DR-1 1927-02-10   622   dyer   rad    7.80
## 4    622  DR-1 1927-02-10   622   dyer   sal    0.09
## 5    734  DR-3 1939-01-07   734   lake   sal    0.05
## 6    734  DR-3 1939-01-07   734     pb   rad    8.41
## 7    734  DR-3 1939-01-07   734     pb  temp  -21.50
## 8    735  DR-3 1930-01-12   735     NA   sal    0.06
## 9    735  DR-3 1930-01-12   735     NA  temp  -26.00
## 10   735  DR-3 1930-01-12   735     pb   rad    7.22
## ..   ...   ...        ...   ...    ...   ...     ...
```

```
## the dplyr syntax
visited_join_survey <- visited %>%
  left_join(survey, by = c('taken' = 'ident'))

explain(visited_join_survey)
```

```
## <SQL>
## SELECT "ident", "site", "dated", "taken", "person", "quant", "reading"
## FROM (SELECT * FROM (SELECT "ident", "site", "dated"
## FROM "Visited") AS "zzz2"
##
## LEFT JOIN
##
## (SELECT "taken", "person", "quant", "reading"
## FROM "Survey") AS "zzz3"
##
## ON ("taken" = "ident")) AS "zzz4"


##

## <PLAN>
##   selectid order from
## 1        1     0    0
```

```
## 2        0     0    0
## 3        0     1    1
##                                                              detail
## 1                                                 SCAN TABLE Survey
## 2                                               SCAN TABLE Visited
## 3 SEARCH SUBQUERY 1 AS zzz3 USING AUTOMATIC COVERING INDEX (taken=?)
```

```
visited_join_survey
```

```
## Source: sqlite 3.8.6 [~/swcarpentry/my_project/data/survey.db]
## From: <derived table> [?? x 7]
##
##    ident  site      dated taken person quant reading
##    (int) (chr)      (chr) (int)  (chr) (chr)   (dbl)
## 1    619  DR-1 1927-02-08   619   dyer   rad    9.82
## 2    619  DR-1 1927-02-08   619   dyer   sal    0.13
## 3    622  DR-1 1927-02-10   622   dyer   rad    7.80
## 4    622  DR-1 1927-02-10   622   dyer   sal    0.09
## 5    734  DR-3 1939-01-07   734   lake   sal    0.05
## 6    734  DR-3 1939-01-07   734     pb   rad    8.41
## 7    734  DR-3 1939-01-07   734     pb  temp  -21.50
## 8    735  DR-3 1930-01-12   735     NA   sal    0.06
## 9    735  DR-3 1930-01-12   735     NA  temp  -26.00
## 10   735  DR-3 1930-01-12   735     pb   rad    7.22
## ..   ...   ...        ...   ...    ...   ...     ...
```

```
x <- collect(visited_join_survey)
```

sqldf

Treats dataframes as database tables.

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Library
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Library not loaded: /opt/X1
##   Referenced from: /Library/Frameworks/R.framework/Resources/modules//R_X11.so
##   Reason: image not found
```

```
## Could not load tcltk.  Will use slower R code instead.
```

```
## Loading required package: RSQLite
```

```
## Loading required package: DBI
```

```
surveydf <- as.data.frame(survey)
visiteddf <- as.data.frame(visited)

sqldf("Select * from visiteddf join surveydf on visiteddf.ident = surveydf.taken")
```

```
##    ident  site      dated taken person quant reading
## 1    619  DR-1 1927-02-08   619   dyer   rad    9.82
## 2    619  DR-1 1927-02-08   619   dyer   sal    0.13
## 3    622  DR-1 1927-02-10   622   dyer   rad    7.80
## 4    622  DR-1 1927-02-10   622   dyer   sal    0.09
## 5    734  DR-3 1939-01-07   734   lake   sal    0.05
## 6    734  DR-3 1939-01-07   734     pb   rad    8.41
## 7    734  DR-3 1939-01-07   734     pb  temp  -21.50
## 8    735  DR-3 1930-01-12   735   <NA>   sal    0.06
## 9    735  DR-3 1930-01-12   735   <NA>  temp  -26.00
## 10   735  DR-3 1930-01-12   735     pb   rad    7.22
## 11   751  DR-3 1930-02-26   751   lake   sal    0.10
## 12   751  DR-3 1930-02-26   751     pb   rad    4.35
## 13   751  DR-3 1930-02-26   751     pb  temp  -18.50
## 14   752  DR-3       <NA>   752   lake   rad    2.19
## 15   752  DR-3       <NA>   752   lake   sal    0.09
## 16   752  DR-3       <NA>   752   lake  temp  -16.00
## 17   752  DR-3       <NA>   752    roe   sal   41.60
## 18   837 MSK-4 1932-01-14   837   lake   rad    1.46
## 19   837 MSK-4 1932-01-14   837   lake   sal    0.21
## 20   837 MSK-4 1932-01-14   837    roe   sal   22.50
## 21   844  DR-1 1932-03-22   844    roe   rad   11.25
```

The RSQLite package

Very powerful. For loading data from a database see `?dbWriteTable`

```
library(RSQLite)

surveydb <- dbConnect(drv = dbDriver("SQLite"), "~/swcarpentry/my_project/data/survey.db")

dbListTables(surveydb)
```

```
## [1] "Person"  "Site"    "Survey"  "Visited"
```

```
dbListFields(surveydb, "visited")
```

```
## [1] "ident" "site"  "dated"
```

```
dbGetQuery(surveydb, "Select * from visited join survey on visited.ident = survey.taken")
```

```
##    ident  site      dated taken person quant reading
## 1    619  DR-1 1927-02-08   619   dyer   rad    9.82
## 2    619  DR-1 1927-02-08   619   dyer   sal    0.13
## 3    622  DR-1 1927-02-10   622   dyer   rad    7.80
```

```
## 4    622  DR-1 1927-02-10   622   dyer   sal    0.09
## 5    734  DR-3 1939-01-07   734   lake   sal    0.05
## 6    734  DR-3 1939-01-07   734     pb   rad    8.41
## 7    734  DR-3 1939-01-07   734     pb  temp  -21.50
## 8    735  DR-3 1930-01-12   735   <NA>   sal    0.06
## 9    735  DR-3 1930-01-12   735   <NA>  temp  -26.00
## 10   735  DR-3 1930-01-12   735     pb   rad    7.22
## 11   751  DR-3 1930-02-26   751   lake   sal    0.10
## 12   751  DR-3 1930-02-26   751     pb   rad    4.35
## 13   751  DR-3 1930-02-26   751     pb  temp  -18.50
## 14   752  DR-3       <NA>   752   lake   rad    2.19
## 15   752  DR-3       <NA>   752   lake   sal    0.09
## 16   752  DR-3       <NA>   752   lake  temp  -16.00
## 17   752  DR-3       <NA>   752    roe   sal   41.60
## 18   837 MSK-4 1932-01-14   837   lake   rad    1.46
## 19   837 MSK-4 1932-01-14   837   lake   sal    0.21
## 20   837 MSK-4 1932-01-14   837    roe   sal   22.50
## 21   844  DR-1 1932-03-22   844    roe   rad   11.25
```