week 7 zrutter

Zai Rutter

2022-05-01

```
#-----------------------------------------
# SQL HW
# Zai rutter
#-----------------------------------------

# We will use occupational employment statistics data
# The file is somewhat large (N>430,000)

# For details, see http://www.bls.gov/oes/current/oes_stru.html

# Peek at the first few rows of the dataset
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, 0x0006): Library not loaded: /
##   Referenced from: /Library/Frameworks/R.framework/Versions/4.1/Resources/modules/R_X11.so
##   Reason: tried: '/opt/X11/lib/libSM.6.dylib' (no such file), '/Library/Frameworks/R.framework/Resou
```

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

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

```
setwd("~/Documents/Upenn/Data 410/Week 7/Homework")
read.table("oesm.csv", sep=";", header = TRUE, fill = TRUE, nrows = 10)
```

```
##       area                                          area_title area_type naics
## 1       99                                                U.S.         1     0
## 2       99                                                U.S.         1     0
## 3  3100003         Northeastern Nebraska nonmetropolitan area         6     0
## 4    42540                        Scranton--Wilkes-Barre, PA         4     0
## 5    25620                                  Hattiesburg, MS         4     0
## 6  3700002          Other North Carolina nonmetropolitan area         6     0
## 7    35620  New York-Northern New Jersey-Long Island, NY-NJ-PA         4     0
## 8  3700003 Western Central North Carolina nonmetropolitan area         6     0
## 9  2900002                North Missouri nonmetropolitan area         6     0
```

```
## 10   23540                                    Gainesville, FL         4     0
##       naics_title own_code occcode                occtitle grouping tot_emp
## 1  Cross-industry     1235 13-2010 Accountants and Auditors    broad 1226910
## 2  Cross-industry     1235 13-2010 Accountants and Auditors    broad 1187310
## 3  Cross-industry     1235 13-2011 Accountants and Auditors   detail     510
## 4  Cross-industry     1235 13-2011 Accountants and Auditors   detail    1830
## 5  Cross-industry     1235 13-2011 Accountants and Auditors   detail     180
## 6  Cross-industry     1235 13-2011 Accountants and Auditors   detail    1270
## 7  Cross-industry     1235 13-2011 Accountants and Auditors   detail  101790
## 8  Cross-industry     1235 13-2011 Accountants and Auditors   detail    1210
## 9  Cross-industry     1235 13-2011 Accountants and Auditors   detail     310
## 10 Cross-industry     1235 13-2011 Accountants and Auditors   detail    1010
##    emp_prse jobs_1000 loc_quotient pct_total h_mean a_mean mean_prse h_pct10
## 1       0.6        NA           NA        NA  36.19  75280       0.3   19.90
## 2       0.6        NA           NA        NA  35.42  73670       0.2   19.64
## 3       6.8     5.445         0.62        NA  29.35  61040       3.0   17.34
## 4       5.3     7.182         0.82        NA  29.28  60900       1.9   17.69
## 5      11.6     3.255         0.37        NA  25.56  53170       5.2   15.74
## 6       7.9     4.274         0.49        NA  36.27  75430       6.2   19.57
## 7       3.6    11.815         1.34        NA  44.79  93160       1.0   24.09
## 8       8.3     4.729         0.54        NA  31.10  64680       4.0   15.25
## 9      13.2     3.508         0.40        NA  28.50  59280       8.3   15.34
## 10      3.0     8.189         0.93        NA  29.10  60530       3.1   17.59
##    h_pct25 h_median h_pct75 h_pct90 a_pct10 a_pct25 a_median a_pct75 a_pct90
## 1    25.04    32.30   43.04   57.18   41400   52090    67190   89520  118930
## 2    24.58    31.70   42.08   55.75   40850   51130    65940   87530  115950
## 3    21.06    26.59   34.19   44.16   36060   43810    55320   71110   91840
## 4    21.73    27.27   34.39   43.23   36790   45200    56730   71530   89910
## 5    18.74    22.38   29.89   40.53   32740   38970    46550   62160   84310
## 6    24.12    30.12   39.28   54.70   40700   50170    62650   81700  113770
## 7    30.07    39.17   52.92   72.88   50100   62550    81480  110080  151580
## 8    21.55    28.08   37.78   49.60   31710   44830    58410   78580  103170
## 9    18.52    22.36   28.12   41.35   31900   38530    46520   58490   86000
## 10   21.76    27.31   33.86   41.04   36600   45260    56810   70420   85370
##    annual hourly year
## 1       0      0 2015
## 2       0      0 2014
## 3       0      0 2014
## 4       0      0 2014
## 5       0      0 2014
## 6       0      0 2014
## 7       0      0 2014
## 8       0      0 2014
## 9       0      0 2014
## 10      0      0 2014
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
# 1. Begin SQL processing by creating a new database called 'conemp'
# Then add a table called 'oesm' (A best practice here is to use an
# if statement to delete a table called 'oesm' if one already exists.)

{

  conemp <- dbConnect(SQLite(), dbname = "oesm.db")

  if (dbExistsTable(conemp, "oesm"))
    dbRemoveTable(conemp, "oesm")


  dbWriteTable(
    conemp, # connection
    "oesm", # new table
    "oesm.csv", # data source
    sep = ";",
    header = TRUE,
    row.names = FALSE
  )
 ## dbDisconnect(conemp)
}

# 2. Display the first 10 rows of all of the data in your table.


data <- dbConnect(SQLite(), dbname = "oesm.db")
q2 <- dbSendQuery(data, "
                    SELECT *
                    FROM oesm")
fetch(q2, n = 10)
```

```
##       area                                          area_title area_type naics
## 1       99                                                U.S.         1     0
## 2       99                                                U.S.         1     0
```

```
## 3  3100003          Northeastern Nebraska nonmetropolitan area         6     0
## 4    42540                        Scranton--Wilkes-Barre, PA         4     0
## 5    25620                                   Hattiesburg, MS         4     0
## 6  3700002           Other North Carolina nonmetropolitan area         6     0
## 7    35620  New York-Northern New Jersey-Long Island, NY-NJ-PA         4     0
## 8  3700003 Western Central North Carolina nonmetropolitan area         6     0
## 9  2900002                North Missouri nonmetropolitan area         6     0
## 10   23540                                   Gainesville, FL         4     0
##       naics_title own_code occcode                occtitle grouping tot_emp
## 1  Cross-industry     1235 13-2010 Accountants and Auditors    broad 1226910
## 2  Cross-industry     1235 13-2010 Accountants and Auditors    broad 1187310
## 3  Cross-industry     1235 13-2011 Accountants and Auditors   detail     510
## 4  Cross-industry     1235 13-2011 Accountants and Auditors   detail    1830
## 5  Cross-industry     1235 13-2011 Accountants and Auditors   detail     180
## 6  Cross-industry     1235 13-2011 Accountants and Auditors   detail    1270
## 7  Cross-industry     1235 13-2011 Accountants and Auditors   detail  101790
## 8  Cross-industry     1235 13-2011 Accountants and Auditors   detail    1210
## 9  Cross-industry     1235 13-2011 Accountants and Auditors   detail     310
## 10 Cross-industry     1235 13-2011 Accountants and Auditors   detail    1010
##    emp_prse jobs_1000 loc_quotient pct_total h_mean a_mean mean_prse h_pct10
## 1        .6        NA           NA            36.19  75280       0.3   19.90
## 2       0.6        NA           NA            35.42  73670       0.2   19.64
## 3       6.8     5.445         0.62            29.35  61040       3.0   17.34
## 4       5.3     7.182         0.82            29.28  60900       1.9   17.69
## 5      11.6     3.255         0.37            25.56  53170       5.2   15.74
## 6       7.9     4.274         0.49            36.27  75430       6.2   19.57
## 7       3.6    11.815         1.34            44.79  93160       1.0   24.09
## 8       8.3     4.729         0.54            31.10  64680       4.0   15.25
## 9      13.2     3.508         0.40            28.50  59280       8.3   15.34
## 10      3.0     8.189         0.93            29.10  60530       3.1   17.59
##    h_pct25 h_median h_pct75 h_pct90 a_pct10 a_pct25 a_median a_pct75 a_pct90
## 1    25.04    32.30   43.04   57.18   41400   52090    67190   89520  118930
## 2    24.58    31.70   42.08   55.75   40850   51130    65940   87530  115950
## 3    21.06    26.59   34.19   44.16   36060   43810    55320   71110   91840
## 4    21.73    27.27   34.39   43.23   36790   45200    56730   71530   89910
## 5    18.74    22.38   29.89   40.53   32740   38970    46550   62160   84310
## 6    24.12    30.12   39.28   54.70   40700   50170    62650   81700  113770
## 7    30.07    39.17   52.92   72.88   50100   62550    81480  110080  151580
## 8    21.55    28.08   37.78   49.60   31710   44830    58410   78580  103170
## 9    18.52    22.36   28.12   41.35   31900   38530    46520   58490   86000
## 10   21.76    27.31   33.86   41.04   36600   45260    56810   70420   85370
##    annual hourly year
## 1       0      0 2015
## 2       0      0 2014
## 3       0      0 2014
## 4       0      0 2014
## 5       0      0 2014
## 6       0      0 2014
## 7       0      0 2014
## 8       0      0 2014
## 9       0      0 2014
## 10      0      0 2014
```

```
dbClearResult(q2)



# 3. Select a subset of columns, 'occtitle', 'h_mean', 'area_title' for
# all of the observations where the grouping variable is 'major.'

q3 <- dbSendQuery(data, "
                  SELECT occtitle,h_mean,area_title
                  FROM oesm
                  GROUP BY 'major'")
# fetch(q3, n = -1)
dbClearResult(q3)


# 4. Display all of the unique values of the variable 'naics'

q4 <- dbSendQuery(data, "
                     SELECT DISTINCT naics
                     FROM oesm")
fetch(q4, n = 10)
```

```
##     naics
## 1       0
## 2       1
## 3      11
## 4  113000
## 5  113300
## 6  115000
## 7  115100
## 8  115200
## 9      21
## 10 211000
```

```
dbClearResult(q4)

# 5. Count all of the observations for each year in the dataset.

q5 <- dbSendQuery(data, "
                  SELECT year, COUNT(*)
                  FROM oesm
                  GROUP BY year")
fetch(q5, n = 10)
```

```
##   year COUNT(*)
## 1 2014   435004
## 2 2015   439942
```

```
dbClearResult(q5)

# 6. Find minimum and maximum of median annual salary (a_mean) for
# each year
```

```
q6 <- dbSendQuery(data, "
                    SELECT year, MIN(a_mean) AS min_salary, MAX(CAST(a_mean AS INT)) AS min_salary
                  FROM oesm
                  WHERE (a_mean IS NOT 'NA')
                  GROUP BY year
                  ")
fetch(q6, n = 10)
```

```
##   year min_salary min_salary
## 1 2014      16600     277420
## 2 2015      16740     286460
```

```
dbClearResult(q6)

# 7. Create a new table from the same database that includes
# the unique values of occcode and occtitle


if(dbExistsTable(data, "newtable")) dbRemoveTable(data, "newtable")

distinct <- dbSendQuery(conemp, "
                        CREATE TABLE newtable AS
                        SELECT DISTINCT occcode,occtitle
                        FROM oesm
                        ")

dbClearResult(distinct)

distinct2 <- dbSendQuery(conemp, "
                          SELECT *
                          FROM newtable
                          ")

fetch(distinct2, n = 10)
```

```
##    occcode                                                    occtitle
## 1  13-2010                                  Accountants and Auditors
## 2  13-2011                                  Accountants and Auditors
## 3  27-2011                                                    Actors
## 4  27-2010                           Actors, Producers, and Directors
## 5  15-2010                                                 Actuaries
## 6  15-2011                                                 Actuaries
## 7  51-9191           Adhesive Bonding Machine Operators and Tenders
## 8  23-1021 Administrative Law Judges, Adjudicators, and Hearing Officers
## 9  11-3010                           Administrative Services Managers
## 10 11-3011                           Administrative Services Managers
```

```
# 8. Using your new SQL skills, pull some information from either table
# and create a visualization (of any kind) with your extracted data. In
# one or two sentences, briefly comment on your visual.

wholedata <- dbSendQuery(conemp, "
```

```
                          SELECT *
                           FROM oesm
                          WHERE (occcode == '15-0000')
                          ")
```

```
## Warning: Closing open result set, pending rows
```

```
# dbClearResult(wholedata)
wholedata2<-fetch(wholedata, n = -1)
```

```
## Warning in result_fetch(res@ptr, n = n): Column `naics`: mixed type, first seen
## values of type integer, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `pct_total`: mixed type, first
## seen values of type string, coercing other values of type integer, real

## Warning in result_fetch(res@ptr, n = n): Column `h_mean`: mixed type, first seen
## values of type real, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `a_mean`: mixed type, first seen
## values of type integer, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `mean_prse`: mixed type, first
## seen values of type real, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `h_pct10`: mixed type, first
## seen values of type real, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `h_pct25`: mixed type, first
## seen values of type real, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `h_median`: mixed type, first
## seen values of type real, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `h_pct75`: mixed type, first
## seen values of type real, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `h_pct90`: mixed type, first
## seen values of type real, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `a_pct10`: mixed type, first
## seen values of type integer, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `a_pct25`: mixed type, first
## seen values of type integer, coercing other values of type string

## Warning in result_fetch(res@ptr, n = n): Column `a_median`: mixed type, first
## seen values of type integer, coercing other values of type string
```

```
## Warning in result_fetch(res@ptr, n = n): Column `a_pct75`: mixed type, first
## seen values of type integer, coercing other values of type string
```

```
## Warning in result_fetch(res@ptr, n = n): Column `a_pct90`: mixed type, first
## seen values of type integer, coercing other values of type string
```

```
ggplot(wholedata2, aes(x=a_mean)) +
  geom_histogram (bins = 100) +
  scale_x_continuous(labels=scales::dollar_format(),
                     breaks = breaks_width(20000))
```