sql_in_r.R

Suchitra

Wed Dec 28 23:44:12 2016

```
install.packages("babynames", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Suchitra/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)
```

```
## package 'babynames' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Suchitra\AppData\Local\Temp\RtmpoZpZvZ\downloaded_packages
```

```
#install.packages("babynames")
library(babynames)
str(babynames)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1825433 obs. of  5 variables:
##  $ year: num  1880 1880 1880 1880 1880 1880 1880 1880 1880 1880 ...
##  $ sex : chr  "F" "F" "F" "F" ...
##  $ name: chr  "Mary" "Anna" "Emma" "Elizabeth" ...
##  $ n   : int  7065 2604 2003 1939 1746 1578 1472 1414 1320 1288 ...
##  $ prop: num  0.0724 0.0267 0.0205 0.0199 0.0179 ...
```

```
install.packages("sqldf", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Suchitra/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)
```

```
## package 'sqldf' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Suchitra\AppData\Local\Temp\RtmpoZpZvZ\downloaded_packages
```

```
#install.packages("sqldf")
library(sqldf)
```

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

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

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

```
sqldf("select count(*) from babynames")
```

```
## Loading required package: tcltk
```

```
## Warning: Quoted identifiers should have class SQL, use DBI::SQL() if the
## caller performs the quoting.
```

```
##   count(*)
## 1  1825433
```

```
sqldf("select * from babynames limit 10")
```

```
##    year sex      name    n       prop
## 1  1880   F      Mary 7065 0.07238359
## 2  1880   F      Anna 2604 0.02667896
## 3  1880   F      Emma 2003 0.02052149
## 4  1880   F Elizabeth 1939 0.01986579
## 5  1880   F    Minnie 1746 0.01788843
## 6  1880   F  Margaret 1578 0.01616720
## 7  1880   F       Ida 1472 0.01508119
## 8  1880   F     Alice 1414 0.01448696
## 9  1880   F    Bertha 1320 0.01352390
## 10 1880   F     Sarah 1288 0.01319605
```

```
sqldf("select year, sex, name from babynames limit 10")
```

```
##    year sex      name
## 1  1880   F      Mary
## 2  1880   F      Anna
## 3  1880   F      Emma
## 4  1880   F Elizabeth
## 5  1880   F    Minnie
## 6  1880   F  Margaret
## 7  1880   F       Ida
## 8  1880   F     Alice
## 9  1880   F    Bertha
## 10 1880   F     Sarah
```

```
sqldf("select year, sex as 'Gender' from babynames limit 10")
```

```
##    year Gender
## 1  1880      F
## 2  1880      F
## 3  1880      F
## 4  1880      F
## 5  1880      F
## 6  1880      F
## 7  1880      F
## 8  1880      F
## 9  1880      F
## 10 1880      F
```

```
#filtering data
sqldf("select year,name, sex as 'Gender' from babynames where sex == 'F' limit 20 ")
```

```
##    year      name Gender
## 1  1880      Mary      F
## 2  1880      Anna      F
## 3  1880      Emma      F
## 4  1880 Elizabeth      F
## 5  1880    Minnie      F
## 6  1880  Margaret      F
## 7  1880       Ida      F
## 8  1880     Alice      F
## 9  1880    Bertha      F
## 10 1880     Sarah      F
## 11 1880     Annie      F
## 12 1880     Clara      F
## 13 1880      Ella      F
## 14 1880  Florence      F
## 15 1880      Cora      F
## 16 1880    Martha      F
## 17 1880     Laura      F
## 18 1880    Nellie      F
## 19 1880     Grace      F
## 20 1880    Carrie      F
```

```
sqldf("select * from babynames where prop > 0.05 limit 20")
```

```
##    year sex    name    n       prop
## 1  1880   F    Mary 7065 0.07238359
## 2  1880   M    John 9655 0.08154561
## 3  1880   M William 9532 0.08050676
## 4  1880   M   James 5927 0.05005912
## 5  1881   F    Mary 6919 0.06999069
## 6  1881   M    John 8769 0.08098149
## 7  1881   M William 8524 0.07871892
## 8  1881   M   James 5442 0.05025673
## 9  1882   F    Mary 8148 0.07042473
## 10 1882   M    John 9557 0.07831488
## 11 1882   M William 9298 0.07619251
## 12 1883   F    Mary 8012 0.06673108
## 13 1883   M    John 8894 0.07907183
## 14 1883   M William 8387 0.07456437
## 15 1884   F    Mary 9217 0.06698985
## 16 1884   M    John 9388 0.07648626
## 17 1884   M William 8897 0.07248597
## 18 1885   F    Mary 9128 0.06430433
## 19 1885   M    John 8756 0.07551726
## 20 1885   M William 8044 0.06937653
```

```
sqldf("select * from babynames where sex!= 'F' limit 10")
```

```
##    year sex    name    n       prop
## 1  1880   M    John 9655 0.08154561
## 2  1880   M William 9532 0.08050676
## 3  1880   M   James 5927 0.05005912
## 4  1880   M Charles 5348 0.04516892
## 5  1880   M  George 5126 0.04329392
## 6  1880   M   Frank 3242 0.02738176
## 7  1880   M  Joseph 2632 0.02222973
## 8  1880   M  Thomas 2534 0.02140203
## 9  1880   M   Henry 2444 0.02064189
## 10 1880   M  Robert 2415 0.02039696
```

```
sqldf("select year,name, 4*prop as 'final_prop' from babynames where prop <= 0.40 limit 10")
```

```
##    year      name final_prop
## 1  1880      Mary 0.28953435
## 2  1880      Anna 0.10671584
## 3  1880      Emma 0.08208596
## 4  1880 Elizabeth 0.07946314
## 5  1880    Minnie 0.07155371
## 6  1880  Margaret 0.06466882
## 7  1880       Ida 0.06032478
## 8  1880     Alice 0.05794785
## 9  1880    Bertha 0.05409559
## 10 1880     Sarah 0.05278418
```

```
#ordering
sqldf("select * from babynames order by year desc limit 20")
```

```
##    year sex      name     n        prop
## 1  2014   F      Emma 20799 0.010729242
## 2  2014   F    Olivia 19674 0.010148906
## 3  2014   F    Sophia 18490 0.009538136
## 4  2014   F  Isabella 16950 0.008743721
## 5  2014   F       Ava 15586 0.008040096
## 6  2014   F       Mia 13442 0.006934106
## 7  2014   F     Emily 12562 0.006480155
## 8  2014   F   Abigail 11985 0.006182507
## 9  2014   F   Madison 10247 0.005285953
## 10 2014   F Charlotte 10048 0.005183298
## 11 2014   F    Harper  9564 0.004933625
## 12 2014   F     Sofia  9542 0.004922276
## 13 2014   F     Avery  9517 0.004909380
## 14 2014   F Elizabeth  9492 0.004896484
## 15 2014   F    Amelia  8727 0.004501856
## 16 2014   F    Evelyn  8692 0.004483801
## 17 2014   F      Ella  8489 0.004379082
## 18 2014   F     Chloe  8469 0.004368765
## 19 2014   F  Victoria  7955 0.004103616
## 20 2014   F    Aubrey  7589 0.003914814
```

```
sqldf("select * from babynames order by year desc,n desc limit 20")# order by 2 conditions
```

```
##    year sex      name     n        prop
## 1  2014   F      Emma 20799 0.010729242
## 2  2014   F    Olivia 19674 0.010148906
## 3  2014   M      Noah 19144 0.009431494
## 4  2014   F    Sophia 18490 0.009538136
## 5  2014   M      Liam 18342 0.009036381
## 6  2014   M     Mason 17092 0.008420555
## 7  2014   F  Isabella 16950 0.008743721
## 8  2014   M     Jacob 16712 0.008233344
## 9  2014   M   William 16687 0.008221027
## 10 2014   M     Ethan 15619 0.007694866
## 11 2014   F       Ava 15586 0.008040096
## 12 2014   M   Michael 15323 0.007549038
## 13 2014   M Alexander 15293 0.007534258
## 14 2014   M     James 14301 0.007045539
## 15 2014   M    Daniel 13829 0.006813003
## 16 2014   M    Elijah 13694 0.006746494
## 17 2014   M  Benjamin 13687 0.006743045
## 18 2014   M     Logan 13579 0.006689838
## 19 2014   F       Mia 13442 0.006934106
## 20 2014   M     Aiden 13296 0.006550415
```

```
sqldf("select * from babynames order by name limit 20")
```

```
##    year sex      name  n         prop
## 1  2007   M     Aaban  5 2.260251e-06
## 2  2009   M     Aaban  6 2.834029e-06
## 3  2010   M     Aaban  9 4.390297e-06
## 4  2011   M     Aaban 11 5.429927e-06
## 5  2012   M     Aaban 11 5.440091e-06
## 6  2013   M     Aaban 14 6.961721e-06
## 7  2014   M     Aaban 16 7.882569e-06
## 8  2011   F     Aabha  7 3.622491e-06
## 9  2012   F     Aabha  5 2.587144e-06
## 10 2014   F     Aabha  9 4.642684e-06
## 11 2003   M     Aabid  5 2.381589e-06
## 12 2008   F Aabriella  5 2.404426e-06
## 13 2014   F Aabriella  5 2.579269e-06
## 14 1987   M     Aadam  5 2.565534e-06
## 15 1988   M     Aadam  5 2.499140e-06
## 16 1993   M     Aadam  7 3.390300e-06
## 17 1994   M     Aadam  6 2.944615e-06
## 18 1995   M     Aadam  6 2.984156e-06
## 19 1996   M     Aadam  5 2.496342e-06
## 20 1997   M     Aadam  5 2.503958e-06
```

```
#condition by string
sqldf("select * from babynames where name like 'Ben%' limit 10")
```

```
##    year sex      name   n         prop
## 1  1880   F    Bennie   8 8.196301e-05
## 2  1880   F      Bena   6 6.147226e-05
## 3  1880   M  Benjamin 490 4.138514e-03
## 4  1880   M       Ben 305 2.576014e-03
## 5  1880   M  Benjiman  28 2.364865e-04
## 6  1880   M    Bennie  26 2.195946e-04
## 7  1880   M   Bennett  15 1.266892e-04
## 8  1880   M Benjamine  13 1.097973e-04
## 9  1880   M  Benjaman  12 1.013514e-04
## 10 1880   M    Benton  10 8.445946e-05
```

```
sqldf("select * from babynames where name like '%man' limit 10")
```

```
##    year sex     name   n         prop
## 1  1880   M   Herman 347 2.930743e-03
## 2  1880   M   Norman 102 8.614865e-04
## 3  1880   M  Sherman  54 4.560811e-04
## 4  1880   M    Lyman  31 2.618243e-04
## 5  1880   M Benjiman  28 2.364865e-04
## 6  1880   M  Freeman  26 2.195946e-04
## 7  1880   M Benjaman  12 1.013514e-04
## 8  1880   M  Coleman  12 1.013514e-04
## 9  1880   M   Truman  12 1.013514e-04
## 10 1880   M  Tillman   9 7.601351e-05
```

```
sqldf("select * from babynames where name like '%man%' limit 10")
```

```
##    year sex     name   n         prop
## 1  1880   F   Amanda 241 2.469136e-03
## 2  1880   F    Mandy  34 3.483428e-04
## 3  1880   F Samantha  21 2.151529e-04
## 4  1880   F    Manda  10 1.024538e-04
## 5  1880   F  Manerva  10 1.024538e-04
## 6  1880   F    Manie  10 1.024538e-04
## 7  1880   F Manervia   5 5.122688e-05
## 8  1880   F  Manuela   5 5.122688e-05
## 9  1880   M   Herman 347 2.930743e-03
## 10 1880   M   Norman 102 8.614865e-04
```

```
sqldf("select * from babynames where name in ('Coleman','Benjamin','Bennie') limit 10")
```

```
##    year sex     name   n         prop
## 1  1880   F   Bennie   8 8.196301e-05
## 2  1880   M Benjamin 490 4.138514e-03
## 3  1880   M   Bennie  26 2.195946e-04
## 4  1880   M  Coleman  12 1.013514e-04
## 5  1881   F   Bennie   9 9.104151e-05
## 6  1881   M Benjamin 481 4.442023e-03
## 7  1881   M   Bennie  32 2.955192e-04
## 8  1881   M  Coleman  16 1.477596e-04
## 9  1882   F   Bennie   7 6.050234e-05
## 10 1882   M Benjamin 478 3.916973e-03
```

```
sqldf("select * from babynames where year between 2000 and 2014 limit 10")
```

```
##    year sex      name     n        prop
## 1  2000   F     Emily 25952 0.013012976
## 2  2000   F    Hannah 23073 0.011569374
## 3  2000   F   Madison 19967 0.010011949
## 4  2000   F    Ashley 17995 0.009023139
## 5  2000   F     Sarah 17687 0.008868700
## 6  2000   F    Alexis 17627 0.008838615
## 7  2000   F  Samantha 17264 0.008656598
## 8  2000   F   Jessica 15704 0.007874375
## 9  2000   F Elizabeth 15088 0.007565497
## 10 2000   F    Taylor 15078 0.007560483
```

```
#multiple filters
sqldf("select * from babynames where year >= 1980 and prop < 0.5 limit 10")
```

```
##    year sex      name     n       prop
## 1  1980   F  Jennifer 58385 0.03279894
## 2  1980   F    Amanda 35820 0.02012260
## 3  1980   F   Jessica 33920 0.01905524
## 4  1980   F   Melissa 31631 0.01776935
## 5  1980   F     Sarah 25741 0.01446052
## 6  1980   F   Heather 19971 0.01121911
## 7  1980   F    Nicole 19916 0.01118821
## 8  1980   F       Amy 19832 0.01114102
## 9  1980   F Elizabeth 19528 0.01097024
## 10 1980   F  Michelle 19120 0.01074104
```

```
sqldf("select * from babynames where year <= 1980 and prop < 0.5 order by prop desc limit 10")
```

```
##    year sex    name    n       prop
## 1  1880   M    John 9655 0.08154561
## 2  1881   M    John 8769 0.08098149
## 3  1880   M William 9532 0.08050676
## 4  1883   M    John 8894 0.07907183
## 5  1881   M William 8524 0.07871892
## 6  1882   M    John 9557 0.07831488
## 7  1884   M    John 9388 0.07648626
## 8  1882   M William 9298 0.07619251
## 9  1886   M    John 9026 0.07582262
## 10 1885   M    John 8756 0.07551726
```

```
#no man in name
sqldf("select * from babynames where name != '%man%' or year >2000 limit 10")
```

```
##    year sex      name    n       prop
## 1  1880   F      Mary 7065 0.07238359
## 2  1880   F      Anna 2604 0.02667896
## 3  1880   F      Emma 2003 0.02052149
## 4  1880   F Elizabeth 1939 0.01986579
## 5  1880   F    Minnie 1746 0.01788843
## 6  1880   F  Margaret 1578 0.01616720
## 7  1880   F       Ida 1472 0.01508119
## 8  1880   F     Alice 1414 0.01448696
## 9  1880   F    Bertha 1320 0.01352390
## 10 1880   F     Sarah 1288 0.01319605
```

```
sqldf("select * from  babynames where prop >0.07 and year not between 2000 and 2014 limit 10")
```

```
##    year sex    name    n       prop
## 1  1880   F    Mary 7065 0.07238359
## 2  1880   M    John 9655 0.08154561
## 3  1880   M William 9532 0.08050676
## 4  1881   M    John 8769 0.08098149
## 5  1881   M William 8524 0.07871892
## 6  1882   F    Mary 8148 0.07042473
## 7  1882   M    John 9557 0.07831488
## 8  1882   M William 9298 0.07619251
## 9  1883   M    John 8894 0.07907183
## 10 1883   M William 8387 0.07456437
```

```
sqldf("select * from babynames where n >10000 order by name desc limit 10")
```

```
##    year sex    name     n        prop
## 1  1985   M Zachary 11341 0.005897117
## 2  1986   M Zachary 11719 0.006102734
## 3  1987   M Zachary 13198 0.006771984
## 4  1988   M Zachary 15864 0.007929272
## 5  1989   M Zachary 18073 0.008626249
## 6  1990   M Zachary 20424 0.009496315
## 7  1991   M Zachary 21382 0.010091023
## 8  1992   M Zachary 24797 0.011817549
## 9  1993   M Zachary 25533 0.012366362
## 10 1994   M Zachary 25132 0.012334010
```

```
#multiple filters- and, or,not
sqldf("select * from babynames where year >= 1980 and prop <0.5 limit 10")
```

```
##    year sex      name     n       prop
## 1  1980   F  Jennifer 58385 0.03279894
## 2  1980   F    Amanda 35820 0.02012260
## 3  1980   F   Jessica 33920 0.01905524
## 4  1980   F   Melissa 31631 0.01776935
## 5  1980   F     Sarah 25741 0.01446052
## 6  1980   F   Heather 19971 0.01121911
## 7  1980   F    Nicole 19916 0.01118821
## 8  1980   F       Amy 19832 0.01114102
## 9  1980   F Elizabeth 19528 0.01097024
## 10 1980   F  Michelle 19120 0.01074104
```

```
sqldf("select * from babynames where year >=1980 and prop <0.5 order by prop desc limit 10")
```

```
##    year sex     name     n       prop
## 1  1980   M  Michael 68666 0.03702727
## 2  1981   M  Michael 68752 0.03692602
## 3  1983   M  Michael 67986 0.03650120
## 4  1982   M  Michael 68204 0.03615491
## 5  1984   M  Michael 67722 0.03610785
## 6  1985   M  Michael 64887 0.03374008
## 7  1986   M  Michael 64184 0.03342417
## 8  1980   F Jennifer 58385 0.03279894
## 9  1987   M  Michael 63631 0.03264950
## 10 1988   M  Michael 64116 0.03204698
```

```
sqldf("select * from babynames where name != '%man%' or year>2000 limit 10")
```

```
##    year sex      name    n       prop
## 1  1880   F      Mary 7065 0.07238359
## 2  1880   F      Anna 2604 0.02667896
## 3  1880   F      Emma 2003 0.02052149
## 4  1880   F Elizabeth 1939 0.01986579
## 5  1880   F    Minnie 1746 0.01788843
## 6  1880   F  Margaret 1578 0.01616720
## 7  1880   F       Ida 1472 0.01508119
## 8  1880   F     Alice 1414 0.01448696
## 9  1880   F    Bertha 1320 0.01352390
## 10 1880   F     Sarah 1288 0.01319605
```

```
sqldf("select * from babynames where prop >0.07 and year not between 2000 and 2014 limit 10")
```

```
##    year sex    name    n       prop
## 1  1880   F    Mary 7065 0.07238359
## 2  1880   M    John 9655 0.08154561
## 3  1880   M William 9532 0.08050676
## 4  1881   M    John 8769 0.08098149
## 5  1881   M William 8524 0.07871892
## 6  1882   F    Mary 8148 0.07042473
## 7  1882   M    John 9557 0.07831488
## 8  1882   M William 9298 0.07619251
## 9  1883   M    John 8894 0.07907183
## 10 1883   M William 8387 0.07456437
```

```
sqldf("select * from babynames where n> 10000 order by name desc limit 10 ")
```

```
##    year sex    name     n        prop
## 1  1985   M Zachary 11341 0.005897117
## 2  1986   M Zachary 11719 0.006102734
## 3  1987   M Zachary 13198 0.006771984
## 4  1988   M Zachary 15864 0.007929272
## 5  1989   M Zachary 18073 0.008626249
## 6  1990   M Zachary 20424 0.009496315
## 7  1991   M Zachary 21382 0.010091023
## 8  1992   M Zachary 24797 0.011817549
## 9  1993   M Zachary 25533 0.012366362
## 10 1994   M Zachary 25132 0.012334010
```

```
#basic aggregations
sqldf("select sum(n) as 'Total_Count' from babynames ")
```

```
##   Total_Count
## 1   337135426
```

```
sqldf("select min(n) as min, max(n) as max from babynames ")
```

```
##   min   max
## 1   5 99680
```

```
sqldf("select year,avg(n) as 'Average' from babynames group by year order by Average desc limit
 10")
```

```
##    year  Average
## 1  1956 363.3259
## 2  1957 363.2612
## 3  1954 363.0194
## 4  1955 361.1035
## 5  1958 358.6455
## 6  1952 356.5439
## 7  1953 355.4379
## 8  1959 353.2381
## 9  1951 351.9237
## 10 1960 348.4117
```

```
sqldf("select year,count(*) as count from babynames group by year limit 10")
```

```
##    year count
## 1  1880  2000
## 2  1881  1935
## 3  1882  2127
## 4  1883  2084
## 5  1884  2297
## 6  1885  2294
## 7  1886  2392
## 8  1887  2373
## 9  1888  2651
## 10 1889  2590
```

```
#count by year
sqldf("select year,n,count(*) as my_count from babynames where n>10000
      group by year order by my_count limit 10")
```

```
##    year     n my_count
## 1  1888 11754        1
## 2  1889 11648        1
## 3  1890 12078        1
## 4  1891 11703        1
## 5  1892 13173        1
## 6  1893 12784        1
## 7  1894 13151        1
## 8  1895 13446        1
## 9  1896 13811        1
## 10 1897 13413        1
```

```
#where command doesn't work on aggregated columns, we use having columns

sqldf("select year,name,sum(n) as 'my_sum' from babynames group by year
      having my_sum >10000 order by my_sum desc limit 10")
```

```
##    year    name  my_sum
## 1  1957  Zonnie 4200026
## 2  1959 Zyndall 4156553
## 3  1960  Zandel 4154810
## 4  1961 Zendell 4139818
## 5  1958  Zettie 4131596
## 6  1956     Zvi 4121206
## 7  1962  Zollie 4035493
## 8  1955 Zephery 4012582
## 9  2007  Zyrese 3991547
## 10 1954  Zeddie 3979056
```

```
#unique
sqldf("select count(distinct name) as 'count_names' from babynames")
```

```
##   count_names
## 1       93889
```

```
#ifelse case
sqldf("select year,n,
      case when year = '1880' then 'Young'
      else 'Old'
      end as 'young_or_old' from babynames limit 10")
```

```
##    year    n young_or_old
## 1  1880 7065        Young
## 2  1880 2604        Young
## 3  1880 2003        Young
## 4  1880 1939        Young
## 5  1880 1746        Young
## 6  1880 1578        Young
## 7  1880 1472        Young
## 8  1880 1414        Young
## 9  1880 1320        Young
## 10 1880 1288        Young
```

```
sqldf("select *,
      case when name != '%man%' then 'Not_a_man'
      when name = 'Ban%' then 'Born_with_Ban'
      else 'Un_Ban_Man' end as 'Name_Fun' from babynames limit 10")
```

```
##    year sex      name    n       prop  Name_Fun
## 1  1880   F      Mary 7065 0.07238359 Not_a_man
## 2  1880   F      Anna 2604 0.02667896 Not_a_man
## 3  1880   F      Emma 2003 0.02052149 Not_a_man
## 4  1880   F Elizabeth 1939 0.01986579 Not_a_man
## 5  1880   F    Minnie 1746 0.01788843 Not_a_man
## 6  1880   F  Margaret 1578 0.01616720 Not_a_man
## 7  1880   F       Ida 1472 0.01508119 Not_a_man
## 8  1880   F     Alice 1414 0.01448696 Not_a_man
## 9  1880   F    Bertha 1320 0.01352390 Not_a_man
## 10 1880   F     Sarah 1288 0.01319605 Not_a_man
```

```
crash = read.csv.sql("crashes.csv",sql ="select * from file")
```

```
## Warning: closing unused connection 5 (crashes.csv)
```

```
roads = read.csv.sql ("roads.csv", sql ="select * from file")
sqldf("select * from crash limit 10")
```

```
##    Year          Road N_Crashes Volume
## 1  1991 Interstate 65        25  40000
## 2  1992 Interstate 65        37  41000
## 3  1993 Interstate 65        45  45000
## 4  1994 Interstate 65        46  45600
## 5  1995 Interstate 65        46  49000
## 6  1996 Interstate 65        59  51000
## 7  1997 Interstate 65        76  52000
## 8  1998 Interstate 65        90  58000
## 9  1999 Interstate 65        95  65000
## 10 2000 Interstate 65        95  74000
```

```
sqldf("select * from roads limit 10")
```

```
##            Road       District Length
## 1 Interstate 65     Greenfield    262
## 2 Interstate 70      Vincennes    156
## 3         US-36 Crawfordsville    139
## 4         US-40     Greenfield    150
## 5         US-52 Crawfordsville    172
```

```
#common column = Road

#joins
#inner join
sqldf("select * from crash join roads on crash.Road = roads.Road limit 10")
```

```
## Warning: closing unused connection 7 (roads.csv)
```

```
## Warning: closing unused connection 5 (roads.csv)
```

```
##    Year          Road N_Crashes Volume          Road   District Length
## 1  1991 Interstate 65        25  40000 Interstate 65 Greenfield    262
## 2  1992 Interstate 65        37  41000 Interstate 65 Greenfield    262
## 3  1993 Interstate 65        45  45000 Interstate 65 Greenfield    262
## 4  1994 Interstate 65        46  45600 Interstate 65 Greenfield    262
## 5  1995 Interstate 65        46  49000 Interstate 65 Greenfield    262
## 6  1996 Interstate 65        59  51000 Interstate 65 Greenfield    262
## 7  1997 Interstate 65        76  52000 Interstate 65 Greenfield    262
## 8  1998 Interstate 65        90  58000 Interstate 65 Greenfield    262
## 9  1999 Interstate 65        95  65000 Interstate 65 Greenfield    262
## 10 2000 Interstate 65        95  74000 Interstate 65 Greenfield    262
```

```
#left join
sqldf("select crash.Year, crash.Volume, roads.* from crash left join roads on
      crash.Road = roads.Road limit 10")
```

```
##    Year Volume          Road   District Length
## 1  1991  40000 Interstate 65 Greenfield    262
## 2  1992  41000 Interstate 65 Greenfield    262
## 3  1993  45000 Interstate 65 Greenfield    262
## 4  1994  45600 Interstate 65 Greenfield    262
## 5  1995  49000 Interstate 65 Greenfield    262
## 6  1996  51000 Interstate 65 Greenfield    262
## 7  1997  52000 Interstate 65 Greenfield    262
## 8  1998  58000 Interstate 65 Greenfield    262
## 9  1999  65000 Interstate 65 Greenfield    262
## 10 2000  74000 Interstate 65 Greenfield    262
```

```
#joining while aggregation
sqldf("select crash.Year, crash.Volume,roads.* from crash left join roads
      on crash.Road = roads.Road order by 1 limit 10")
```

```
##    Year Volume          Road       District Length
## 1  1991  40000 Interstate 65     Greenfield    262
## 2  1991  76000 Interstate 70      Vincennes    156
## 3  1991  21000         US-40     Greenfield    150
## 4  1991  35200         US-36 Crawfordsville    139
## 5  1991  20350          <NA>           <NA>     NA
## 6  1992  41000 Interstate 65     Greenfield    262
## 7  1992  79000 Interstate 70      Vincennes    156
## 8  1992  21500         US-40     Greenfield    150
## 9  1992  45000         US-36 Crawfordsville    139
## 10 1992  21200          <NA>           <NA>     NA
```

```
sqldf("select crash.Year, crash.Volume, roads.* from crash left join roads on
      crash.Road  = roads.Road where roads.Road != 'US-36'  order by 1 limit 10")
```

```
##    Year Volume          Road   District Length
## 1  1991  40000 Interstate 65 Greenfield    262
## 2  1991  76000 Interstate 70  Vincennes    156
## 3  1991  21000         US-40 Greenfield    150
## 4  1992  41000 Interstate 65 Greenfield    262
## 5  1992  79000 Interstate 70  Vincennes    156
## 6  1992  21500         US-40 Greenfield    150
## 7  1993  45000 Interstate 65 Greenfield    262
## 8  1993  82000 Interstate 70  Vincennes    156
## 9  1993  23000         US-40 Greenfield    150
## 10 1994  45600 Interstate 65 Greenfield    262
```

```
sqldf("select Road, avg(roads.Length) as 'Avg_Length',
      avg(N_Crashes) as 'Avg_Crash' from roads join crash using(Road) group by Road limit 10")
```

```
##            Road Avg_Length Avg_Crash
## 1 Interstate 65        262 107.81818
## 2 Interstate 70        156  65.18182
## 3         US-36        139  48.00000
## 4         US-40        150  68.68182
```

```
roads$Year <- crash$Year[1:5]
sqldf("select crash.Year, crash.Volume, roads.* from crash left join roads
      on crash.Road = roads.Road  and crash.Year = roads.Year order by 1 limit 10")
```

```
##    Year Volume          Road   District Length Year
## 1  1991  40000 Interstate 65 Greenfield    262 1991
## 2  1991  76000          <NA>       <NA>     NA   NA
## 3  1991  21000          <NA>       <NA>     NA   NA
## 4  1991  35200          <NA>       <NA>     NA   NA
## 5  1991  20350          <NA>       <NA>     NA   NA
## 6  1992  41000          <NA>       <NA>     NA   NA
## 7  1992  79000 Interstate 70  Vincennes    156 1992
## 8  1992  21500          <NA>       <NA>     NA   NA
## 9  1992  45000          <NA>       <NA>     NA   NA
## 10 1992  21200          <NA>       <NA>     NA   NA
```

```
#The string functions in sqldf package are implemented under different function names;
#i.e., you can't use the left command to extract characters from the left.

library(RSQLite)
help("initExtension")
```

```
## starting httpd help server ...
```

```
##  done
```

```
#string functions
sqldf("select name, leftstr(name,3) as 'First_3' from babynames
      order by First_3 desc limit 10")
```

```
##         name First_3
## 1      Zzyzx     Zzy
## 2    Zyyanna     Zyy
## 3      Zyyon     Zyy
## 4     Zyvion     Zyv
## 5  Zytavious     Zyt
## 6  Zytavious     Zyt
## 7  Zytavious     Zyt
## 8  Zytavious     Zyt
## 9  Zytavious     Zyt
## 10 Zytavious     Zyt
```

```
sqldf("select name, reverse(name) as 'Rev_Name' from babynames limit 10")
```

```
##         name  Rev_Name
## 1       Mary      yraM
## 2       Anna      annA
## 3       Emma      ammE
## 4  Elizabeth htebazilE
## 5     Minnie    einniM
## 6   Margaret  teragraM
## 7        Ida       adI
## 8      Alice     ecilA
## 9     Bertha    ahtreB
## 10     Sarah     haraS
```

```
sqldf("select name, rightstr(name,3) as 'Back_3' from babynames
      order by Back_3 desc limit 10")
```

```
##     name Back_3
## 1  Lizzy    zzy
## 2  Lizzy    zzy
## 3   Izzy    zzy
## 4  Lizzy    zzy
## 5   Izzy    zzy
## 6  Lizzy    zzy
## 7  Lizzy    zzy
## 8   Izzy    zzy
## 9  Lizzy    zzy
## 10 Lizzy    zzy
```