Illustrating SQL using R

The “sqldf” package can be used to run SQL queries in R.

Here is an illustration of some SQL commands using a morbid but well-known data-set recording the survival (or not) of passengers aboard the Titanic.

Check the first few entries of the data set to get an idea of its structure and content:

```
sqldf("select * from titanic3 limit 5")
```

```
##   pclass survived                             name    sex     age sibsp
## 1    1st        1   Allen, Miss. Elisabeth Walton female 29.0000     0
## 2    1st        1  Allison, Master. Hudson Trevor   male  0.9167     1
## 3    1st        0    Allison, Miss. Helen Loraine female  2.0000     1
## 4    1st        0 Allison, Mr. Hudson Joshua Crei   male 30.0000     1
## 5    1st        0 Allison, Mrs. Hudson J C (Bessi female 25.0000     1
##   parch ticket     fare   cabin    embarked boat body
## 1     0  24160 211.3375      B5 Southampton    2   NA
## 2     2 113781 151.5500 C22 C26 Southampton   11   NA
## 3     2 113781 151.5500 C22 C26 Southampton        NA
## 4     2 113781 151.5500 C22 C26 Southampton       135
## 5     2 113781 151.5500 C22 C26 Southampton        NA
##                         homedest
## 1                     St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
```

How many passengers are there?

```
sqldf("select count(*) from titanic3")
```

```
##   count(*)
## 1     1309
```

How many passengers are there with ages listed?

```
sqldf("select count(age) from titanic3")
```

```
##   count(age)
## 1       1046
```

How many passengers 10 or younger survived (of those with ages listed)?

```
sqldf("select count(*) from titanic3
      where age<=10 and survived=1")
```

```
##   count(*)
## 1       50
```

Who were they?

```
sqldf("select name,age from titanic3
      where age <=10 and survived =1
      limit 10")
```

```
##                                 name    age
## 1     Allison, Master. Hudson Trevor 0.9167
## 2          Dodge, Master. Washington 4.0000
## 3    Spedden, Master. Robert Douglas 6.0000
## 4          Becker, Master. Richard F 1.0000
## 5        Becker, Miss. Marion Louise 4.0000
## 6      Caldwell, Master. Alden Gates 0.8333
## 7  Collyer, Miss. Marjorie \\"Lottie 8.0000
## 8     Davies, Master. John Morgan Jr 8.0000
## 9      Drew, Master. Marshall Brines 8.0000
## 10         Hamalainen, Master. Viljo 0.6667
```

How many distinct cabins contained passengers?

```
sqldf("select count(distinct cabin) from titanic3")
```

```
##   count(distinct cabin)
## 1                   187
```

Who were the oldest survivors of the wreck?

```
sqldf("select name,age from titanic3
      where (survived=1 and age is not null)
      order by age desc
      limit 10")
```

```
##                              name age
## 1  Barkworth, Mr. Algernon Henry W  80
## 2   Cavendish, Mrs. Tyrell William  76
## 3   Compton, Mrs. Alexander Taylor  64
## 4  Crosby, Mrs. Edward Gifford (Ca  64
## 5  Andrews, Miss. Kornelia Theodos  63
## 6           Turkula, Mrs. (Hedwig)  63
## 7  Stone, Mrs. George Nelson (Mart  62
## 8               Harris, Mr. George  62
## 9  Bucknell, Mrs. William Robert (  60
## 10 Fortune, Mrs. Mark (Mary McDoug  60
```

How many passengers traveling to Indianapolis were between the ages of 18 and 21?

```
sqldf("select age,embarked from titanic3
      where age between 18 and 21
      and homedest=\"Indianapolis, IN\"")
```

```
## [1] age      embarked
## <0 rows> (or 0-length row.names)
```

What were the embarkation points for passengers aboard the Titanic?

```
sqldf("select distinct embarked from titanic3")
```

```
##      embarked
## 1 Southampton
## 2   Cherbourg
## 3
## 4  Queenstown
```

Which enigmatic passengers have " " listed as their embarkation point?

```
sqldf("select name,age,embarked,homedest from titanic3 where embarked
      not in (\"Queenstown\",\"Cherbourg\",\"Southampton\")")
```

```
##                              name age embarked       homedest
## 1             Icard, Miss. Amelie  38
## 2 Stone, Mrs. George Nelson (Mart  62          Cincinatti, OH
```

How many passengers embarked from the British Isles?

```
sqldf("select count(*) from titanic3
      where embarked in(\"Queenstown\",\"Southampton\")
      limit 10")
```

```
##   count(*)
## 1     1037
```

What was the survival rate for those who paid the lowest fares (less than one fifth of the average) versus of those who paid the highest fares (more than five times the average)?

First, how many passengers paid the lowest fares?

```
sqldf("select count(fare) from titanic3
      where fare<.2*(select avg(fare) from titanic3)")
```

```
##   count(fare)
## 1          28
```

How man of those paying the **lowest** fare survived?

```
sqldf("select count(*)from titanic3
      where fare<.2*(select avg(fare) from titanic3)
      and survived=1")
```

```
##   count(*)
## 1        3
```

How many passengers paid the highest fares?

```
sqldf("select count(*) from titanic3
      where fare>5*(select avg(fare) from titanic3)")
```

```
##   count(*)
## 1       38
```

And how many of those paying the **highest** fares survived?

```
sqldf("select count(*) from titanic3
      where fare>5*(select avg(fare) from titanic3)
      and survived=1")
```

```
##   count(*)
## 1       26
```

Which cabin names started with ‘A’?

```
sqldf("select distinct cabin from titanic3
      where cabin like \"A%\"")
```

```
##    cabin
## 1    A36
## 2    A23
## 3    A31
## 4    A21
## 5     A9
## 6    A14
## 7    A34
## 8    A16
## 9    A20
## 10   A18
## 11   A29
## 12    A5
## 13   A24
## 14   A32
## 15   A11
## 16   A10
## 17   A26
## 18    A6
## 19    A7
## 20   A19
```

Many passengers are listed as assigned to more than one cabin. Which cabin names were combined in this way? To answer this, note that single cabin names are always four characters at most (a letter plus up to three numerals). So, which cabin assignments have five or more characters? This should give us all the cabins that are listed together.

```
sqldf("select distinct cabin from titanic3
      where cabin like \"_____%\"")
```

```
##              cabin
## 1          C22 C26
## 2          C62 C64
## 3          B58 B60
## 4      B51 B53 B55
## 5          B96 B98
## 6      C23 C25 C27
## 7          D10 D12
## 8          B82 B84
## 9      B52 B54 B56
## 10 B57 B59 B63 B66
## 11         C55 C57
## 12         E39 E41
## 13           F G63
## 14           F E57
## 15           F E46
## 16           F G73
## 17           F E69
```