

Introduction to SQL

Greg Ridgeway

Ruth Moyer

Li Sian Goh

2025-11-30

Table of contents

1	Introduction	1
2	Getting the data into proper form	2
3	Setting up the database	6
4	SQL queries (SELECT, WHERE, FROM)	7
4.1	Exercises	11
5	GROUP BY and aggregation functions	12
5.1	Exercises	13
6	ORDER BY and UPDATE	14
6.1	Exercises	19
7	Solutions to the exercises	20

1 Introduction

Some datasets are far too large for R to handle by itself. Structured Query Language (“SQL”) is a widely used international standard language for managing data stored in a relational database management system (RDMS). A relational database management system itself is an approach to managing data using a structure that can be contrasted against the “flat file” approach we have been using thus far with R. Why use SQL? R does not work very well with really huge datasets. A relational database management system offers a way of storing large amounts of information more efficiently and reducing the size of the dataset that we are working with. There are numerous relational database management systems such as Oracle DBMS, Microsoft Access, Microsoft SQL Server, PostgreSQL, and MySQL. We are going to

use [SQLite](#), which is probably the most widely deployed database system. SQLite is in your phone, car, airplanes, thermostats, and numerous appliances. We are going to hook up SQLite to R so that R can handle large datasets.

These are some basic clauses in a SQL query that we will explore:

- SELECT fields or functions of fields
- FROM tables queried
- WHERE conditions for selecting a record
- GROUP BY list of fields to group
- ORDER BY list of fields to sort by

However, before being able to use SQL as a tool in R, we first need to load the [RSQLite](#) package, which provides the software tools to connect to a SQLite database.

```
library(dplyr)
library(RSQLite)
```

2 Getting the data into proper form

We will be working with Chicago crime data, which is accessible in comma-separated value (csv) format. Before we can even begin learning SQL, we are going to have to do a fair bit of work to acquire the dataset, format it so that it is ready for SQLite, and then load it into the SQLite database.

Navigate to the Chicago open data website to get the [data](#). Click the “Export” button and select the “CSV” option, or directly download from [here](#)

The Chicago crime data is huge, more than 2.0 Gb. It contains over 8.3 million records on all crimes reported to the Chicago police department since 2001. R does not handle really large datasets well. By using SQL, you will learn how to more efficiently work with large datasets and learn a data language that is used absolutely everywhere.

Let's use `scan()` to just peek at the first five rows of the file.

```
scan(what="", file="Crimes_-_2001_to_present.csv", nlines=5, sep="\n")
```

```
[1] "ID,Case Number,Date,Block,IUCR,Primary Type,Description,Location Description,Arrest,Domestic,Beat,Community Area,Year,Month,Day,Hour,Minute,Seconds,Year2,Month2,Day2,Hour2,Minute2,Seconds2,Block2,Block3,Block4,Block5,Block6,Block7,Block8,Block9,Block10,Block11,Block12,Block13,Block14,Block15,Block16,Block17,Block18,Block19,Block20,Block21,Block22,Block23,Block24,Block25,Block26,Block27,Block28,Block29,Block30,Block31,Block32,Block33,Block34,Block35,Block36,Block37,Block38,Block39,Block40,Block41,Block42,Block43,Block44,Block45,Block46,Block47,Block48,Block49,Block50,Block51,Block52,Block53,Block54,Block55,Block56,Block57,Block58,Block59,Block500,Block501,Block502,Block503,Block504,Block505,Block506,Block507,Block508,Block509,Block510,Block511,Block512,Block513,Block514,Block515,Block516,Block517,Block518,Block519,Block520,Block521,Block522,Block523,Block524,Block525,Block526,Block527,Block528,Block529,Block5200,Block5201,Block5202,Block5203,Block5204,Block5205,Block5206,Block5207,Block5208,Block5209,Block5210,Block5211,Block5212,Block5213,Block5214,Block5215,Block5216,Block5217,Block5218,Block5219,Block5220,Block5221,Block5222,Block5223,Block5224,Block5225,Block5226,Block5227,Block5228,Block5229,Block5230,Block5231,Block5232,Block5233,Block5234,Block5235,Block5236,Block5237,Block5238,Block5239,Block5240,Block5241,Block5242,Block5243,Block5244,Block5245,Block5246,Block5247,Block5248,Block5249,Block5250,Block5251,Block5252,Block5253,Block5254,Block5255,Block5256,Block5257,Block5258,Block5259,Block5260,Block5261,Block5262,Block5263,Block5264,Block5265,Block5266,Block5267,Block5268,Block5269,Block5270,Block5271,Block5272,Block5273,Block5274,Block5275,Block5276,Block5277,Block5278,Block5279,Block5280,Block5281,Block5282,Block5283,Block5284,Block5285,Block5286,Block5287,Block5288,Block5289,Block5290,Block5291,Block5292,Block5293,Block5294,Block5295,Block5296,Block5297,Block5298,Block5299,Block52000,Block52001,Block52002,Block52003,Block52004,Block52005,Block52006,Block52007,Block52008,Block52009,Block52010,Block52011,Block52012,Block52013,Block52014,Block52015,Block52016,Block52017,Block52018,Block52019,Block52020,Block52021,Block52022,Block52023,Block52024,Block52025,Block52026,Block52027,Block52028,Block52029,Block52030,Block52031,Block52032,Block52033,Block52034,Block52035,Block52036,Block52037,Block52038,Block52039,Block52040,Block52041,Block52042,Block52043,Block52044,Block52045,Block52046,Block52047,Block52048,Block52049,Block52050,Block52051,Block52052,Block52053,Block52054,Block52055,Block52056,Block52057,Block52058,Block52059,Block52060,Block52061,Block52062,Block52063,Block52064,Block52065,Block52066,Block52067,Block52068,Block52069,Block52070,Block52071,Block52072,Block52073,Block52074,Block52075,Block52076,Block52077,Block52078,Block52079,Block52080,Block52081,Block52082,Block52083,Block52084,Block52085,Block52086,Block52087,Block52088,Block52089,Block52090,Block52091,Block52092,Block52093,Block52094,Block52095,Block52096,Block52097,Block52098,Block52099,Block520000,Block520001,Block520002,Block520003,Block520004,Block520005,Block520006,Block520007,Block520008,Block520009,Block520010,Block520011,Block520012,Block520013,Block520014,Block520015,Block520016,Block520017,Block520018,Block520019,Block520020,Block520021,Block520022,Block520023,Block520024,Block520025,Block520026,Block520027,Block520028,Block520029,Block520030,Block520031,Block520032,Block520033,Block520034,Block520035,Block520036,Block520037,Block520038,Block520039,Block520040,Block520041,Block520042,Block520043,Block520044,Block520045,Block520046,Block520047,Block520048,Block520049,Block520050,Block520051,Block520052,Block520053,Block520054,Block520055,Block520056,Block520057,Block520058,Block520059,Block520060,Block520061,Block520062,Block520063,Block520064,Block520065,Block520066,Block520067,Block520068,Block520069,Block520070,Block520071,Block520072,Block520073,Block520074,Block520075,Block520076,Block520077,Block520078,Block520079,Block520080,Block520081,Block520082,Block520083,Block520084,Block520085,Block520086,Block520087,Block520088,Block520089,Block520090,Block520091,Block520092,Block520093,Block520094,Block520095,Block520096,Block520097,Block520098,Block520099,Block5200000,Block5200001,Block5200002,Block5200003,Block5200004,Block5200005,Block5200006,Block5200007,Block5200008,Block5200009,Block5200010,Block5200011,Block5200012,Block5200013,Block5200014,Block5200015,Block5200016,Block5200017,Block5200018,Block5200019,Block5200020,Block5200021,Block5200022,Block5200023,Block5200024,Block5200025,Block5200026,Block5200027,Block5200028,Block5200029,Block5200030,Block5200031,Block5200032,Block5200033,Block5200034,Block5200035,Block5200036,Block5200037,Block5200038,Block5200039,Block5200040,Block5200041,Block5200042,Block5200043,Block5200044,Block5200045,Block5200046,Block5200047,Block5200048,Block5200049,Block5200050,Block5200051,Block5200052,Block5200053,Block5200054,Block5200055,Block5200056,Block5200057,Block5200058,Block5200059,Block5200060,Block5200061,Block5200062,Block5200063,Block5200064,Block5200065,Block5200066,Block5200067,Block5200068,Block5200069,Block5200070,Block5200071,Block5200072,Block5200073,Block5200074,Block5200075,Block5200076,Block5200077,Block5200078,Block5200079,Block5200080,Block5200081,Block5200082,Block5200083,Block5200084,Block5200085,Block5200086,Block5200087,Block5200088,Block5200089,Block52000800,Block52000801,Block52000802,Block52000803,Block52000804,Block52000805,Block52000806,Block52000807,Block52000808,Block52000809,Block520008010,Block520008011,Block520008012,Block520008013,Block520008014,Block520008015,Block520008016,Block520008017,Block520008018,Block520008019,Block520008020,Block520008021,Block520008022,Block520008023,Block520008024,Block520008025,Block520008026,Block520008027,Block520008028,Block520008029,Block520008030,Block520008031,Block520008032,Block520008033,Block520008034,Block520008035,Block520008036,Block520008037,Block520008038,Block520008039,Block520008040,Block520008041,Block520008042,Block520008043,Block520008044,Block520008045,Block520008046,Block520008047,Block520008048,Block520008049,Block520008050,Block520008051,Block520008052,Block520008053,Block520008054,Block520008055,Block520008056,Block520008057,Block520008058,Block520008059,Block520008060,Block520008061,Block520008062,Block520008063,Block520008064,Block520008065,Block520008066,Block520008067,Block520008068,Block520008069,Block520008070,Block520008071,Block520008072,Block520008073,Block520008074,Block520008075,Block520008076,Block520008077,Block520008078,Block520008079,Block520008080,Block520008081,Block520008082,Block520008083,Block520008084,Block520008085,Block520008086,Block520008087,Block520008088,Block520008089,Block520008090,Block520008091,Block520008092,Block520008093,Block520008094,Block520008095,Block520008096,Block520008097,Block520008098,Block520008099,Block520008000,Block520008001,Block520008002,Block520008003,Block520008004,Block520008005,Block520008006,Block520008007,Block520008008,Block520008009,Block520008010,Block520008011,Block520008012,Block520008013,Block520008014,Block520008015,Block520008016,Block520008017,Block520008018,Block520008019,Block520008020,Block520008021,Block520008022,Block520008023,Block520008024,Block520008025,Block520008026,Block520008027,Block520008028,Block520008029,Block520008030,Block520008031,Block520008032,Block520008033,Block520008034,Block520008035,Block520008036,Block520008037,Block520008038,Block520008039,Block520008040,Block520008041,Block520008042,Block520008043,Block520008044,Block520008045,Block520008046,Block520008047,Block520008048,Block520008049,Block520008050,Block520008051,Block520008052,Block520008053,Block520008054,Block520008055,Block520008056,Block520008057,Block520008058,Block520008059,Block520008060,Block520008061,Block520008062,Block520008063,Block520008064,Block520008065,Block520008066,Block520008067,Block520008068,Block520008069,Block520008070,Block520008071,Block520008072,Block520008073,Block520008074,Block520008075,Block520008076,Block520008077,Block520008078,Block520008079,Block520008080,Block520008081,Block520008082,Block520008083,Block520008084,Block520008085,Block520008086,Block520008087,Block520008088,Block520008089,Block520008090,Block520008091,Block520008092,Block520008093,Block520008094,Block520008095,Block520008096,Block520008097,Block520008098,Block520008099,Block5200080000,Block5200080001,Block5200080002,Block5200080003,Block5200080004,Block5200080005,Block5200080006,Block5200080007,Block5200080008,Block5200080009,Block5200080010,Block5200080011,Block5200080012,Block5200080013,Block5200080014,Block5200080015,Block5200080016,Block5200080017,Block5200080018,Block5200080019,Block5200080020,Block5200080021,Block5200080022,Block5200080023,Block5200080024,Block5200080025,Block5200080026,Block5200080027,Block5200080028,Block5200080029,Block5200080030,Block5200080031,Block5200080032,Block5200080033,Block5200080034,Block5200080035,Block5200080036,Block5200080037,Block5200080038,Block5200080039,Block5200080040,Block5200080041,Block5200080042,Block5200080043,Block5200080044,Block5200080045,Block5200080046,Block5200080047,Block5200080048,Block5200080049,Block5200080050,Block5200080051,Block5200080052,Block5200080053,Block5200080054,Block5200080055,Block5200080056,Block5200080057,Block5200080058,Block5200080059,Block5200080060,Block5200080061,Block5200080062,Block5200080063,Block5200080064,Block5200080065,Block5200080066,Block5200080067,Block5200080068,Block5200080069,Block5200080070,Block5200080071,Block5200080072,Block5200080073,Block5200080074,Block5200080075,Block5200080076,Block5200080077,Block5200080078,Block5200080079,Block5200080080,Block5200080081,Block5200080082,Block5200080083,Block5200080084,Block5200080085,Block5200080086,Block5200080087,Block5200080088,Block5200080089,Block5200080090,Block5200080091,Block5200080092,Block5200080093,Block5200080094,Block5200080095,Block5200080096,Block5200080097,Block5200080098,Block5200080099,Block5200080100,Block5200080101,Block5200080102,Block5200080103,Block5200080104,Block5200080105,Block5200080106,Block5200080107,Block5200080108,Block5200080109,Block5200080110,Block5200080111,Block5200080112,Block5200080113,Block5200080114,Block5200080115,Block5200080116,Block5200080117,Block5200080118,Block5200080119,Block5200080120,Block5200080121,Block5200080122,Block5200080123,Block5200080124,Block5200080125,Block5200080126,Block5200080127,Block5200080128,Block5200080129,Block5200080130,Block5200080131,Block5200080132,Block5200080133,Block5200080134,Block5200080135,Block5200080136,Block5200080137,Block5200080138,Block5200080139,Block5200080140,Block5200080141,Block5200080142,Block5200080143,Block5200080144,Block5200080145,Block5200080146,Block5200080147,Block5200080148,Block5200080149,Block5200080150,Block5200080151,Block5200080152,Block5200080153,Block5200080154,Block5200080155,Block5200080156,Block5200080157,Block5200080158,Block5200080159,Block5200080160,Block5200080161,Block5200080162,Block5200080163,Block5200080164,Block5200080165,Block5200080166,Block5200080167,Block5200080168,Block5200080169,Block5200080170,Block5200080171,Block5200080172,Block5200080173,Block5200080174,Block5200080175,Block5200080176,Block5200080177,Block5200080178,Block5200080179,Block5200080180,Block5200080181,Block5200080182,Block5200080183,Block5200080184,Block5200080185,Block5200080186,Block5200080187,Block5200080188,Block5200080189,Block5200080190,Block5200080191,Block5200080192,Block5200080193,Block5200080194,Block5200080195,Block5200080196,Block5200080197,Block5200080198,Block5200080199,Block5200080200,Block5200080201,Block5200080202,Block5200080203,Block5200080204,Block5200080205,Block5200080206,Block5200080207,Block5200080208,Block5200080209,Block5200080210,Block5200080211,Block5200080212,Block5200080213,Block5200080214,Block5200080215,Block5200080216,Block5200080217,Block5200080218,Block5200080219,Block5200080220,Block5200080221,Block5200080222,Block5200080223,Block5200080224,Block5200080225,Block5200080226,Block5200080227,Block5200080228,Block5200080229,Block5200080230,Block5200080231,Block5200080232,Block5200080233,Block5200080234,Block5200080235,Block5200080236,Block5200080237,Block5200080238,Block5200080239,Block5200080240,Block5200080241,Block5200080242,Block5200080243,Block5200080244,Block5200080245,Block5200080246,Block5200080247,Block5200080248,Block5200080249,Block5200080250,Block5200080251,Block5200080252,Block5200080253,Block5200080254,Block5200080255,Block5200080256,Block5200080257,Block5200080258,Block5200080259,Block5200080260,Block5200080261,Block5200080262,Block5200080263,Block5200080264,Block5200080265,Block5200080266,Block5200080267,Block5200080268,Block5200080269,Block5200080270,Block5200080271,Block5200080272,Block5200080273,Block5200080274,Block5200080275,Block5200080276,Block5200080277,Block5200080278,Block5200080279,Block5200080280,Block5200080281,Block5200080282,Block5200080283,Block5200080284,Block5200080285,Block5200080286,Block5200080287,Block5200080288,Block5200080289,Block5200080290,Block5200080291,Block5200080292,Block5200080293,Block5200080294,Block5200080295,Block5200080296,Block5200080297,Block5200080298,Block5200080299,Block5200080300,Block5200080301,Block5200080302,Block5200080303,Block5200080304,Block5200080305,Block5200080306,Block5200080307,Block5200080308,Block5200080309,Block5200080310,Block5200080311,Block5200080312,Block5200080313,Block5200080314,Block5200080315,Block5200080316,Block5200080317,Block5200080318,Block5200080319,Block5200080320,Block5200080321,Block5200080322,Block5200080323,Block5200080324,Block5200080325,Block5200080326,Block5200080327,Block5200080328,Block5200080329,Block5200080330,Block5200080331,Block5200080332,Block5200080333,Block5200080334,Block5200080335,Block5200080336,Block5200080337,Block5200080338,Block5200080339,Block5200080340,Block5200080341,Block5200080342,Block5200080343,Block5200080344,Block5200080345,Block5200080346,Block5200080347,Block5200080348,Block5200080349,Block5200080350,Block5200080351,Block5200080352,Block5200080353,Block5200080354,Block5200080355,Block5200080356,Block5200080357,Block5200080358,Block5200080359,Block5200080360,Block5200080361,Block5200080362,Block5200080363,Block5200080364,Block5200080365,Block5200080366,Block5200080367,Block5200080368,Block5200080369,Block5200080370,Block5200080371,Block5200080372,Block5200080373,Block5200080374,Block5200080375,Block5200080376,Block5200080377,Block5200080378,Block5200080379,Block5200080380,Block5200080381,Block5200080382,Block5200080383,Block5200080384,Block5200080385,Block5200080386,Block5200080387,Block5200080388,Block5200080389,Block5200080390,Block5200080391,Block5200080392,Block5200080393,Block5200080394,Block5200080395,Block5200080396,Block5200080397,Block5200080398,Block5200080399,Block5200080400,Block5200080401,Block5200080402,Block5200080403,Block5200080404,Block5200080405,Block5200080406,Block5200080407,Block5200080408,Block5200080409,Block5200080410,Block5200080411,Block5200080412,Block5200080413,Block5200080414,Block5200080415,Block5200080416,Block5200080417,Block5200080418,Block5200080419,Block5200080
```

`scan()` is a very basic R function that reads in plain text files. We have told it to read in text (`what=""`), the name of the file, to only read in 5 lines (`nlines=5`), and to start a new row whenever it reaches a line feed character (`sep="\n"`). Using `scan()` without `nlines=5` would cause R to try to read in the whole dataset and that could take a lot of time and you might run out of memory.

You can see that the first row contains the column names. The second row contains the first reported crime in the file. You can see date and time, address, crime descriptions, longitude and latitude of the crime, and other information.

Let's try to load this file into a SQLite database. There are two steps. First, using `dbConnect()` we need to tell R to make a connection to a new SQLite database that we will call `chicagocrime.db`. This will be a file in your working folder that SQLite will use to store the data.

```
# create a connection to the database
con <- dbConnect(SQLite(), dbname="chicagocrime.db")
```

Then using `dbWriteTable()` we tell R to read in the csv file and store its contents in a new table in the database. We will call that new table `crime`. Make sure that your path is set to the correct folder where you want the database to be stored.

```
# write a table called "crime" into the SQLite database
dbWriteTable(con,
             "crime", # the new table in the database
             "Crimes_-_2001_to_present.csv",
             row.names=FALSE,
             header=TRUE)      # first row has column names
```

```
Error in connection_import_file(conn@ptr, name, value, sep, eol, skip): RS_sqlite_import: Cr
```

Looks like there is a problem with the dataset. SQLite was expecting 22 columns, but row 4 had 23. Notice from when we ran `scan()` earlier, the fourth row has a "(41.908417822, -87.67740693)". SQLite thinks that these two numbers belong in two different columns instead of a single `Location` column.

SQLite is very particular about the formatting of a file. It can easily read in a csv file, but this dataset has some commas in places that confuse SQLite. For example, there is a row in this file that looks like this:

```
[1] "10000153, HY189345, 03/18/2015 12:20:00 PM, 091XX S UNIVERSITY AVE, 0483, BATTERY, AGG PRO. EM
```

You see that the location description for this crime is "SCHOOL, PUBLIC, BUILDING". Those commas inside the quotes are going to cause SQLite problems. SQLite is going to think that SCHOOL, PUBLIC, and BUILDING are all separate columns rather than in one column describing the location.

To resolve this, we are going to change all the commas that separate the columns into something else besides commas, leaving the commas in elements like "SCHOOL, PUBLIC, BUILDING" alone. It does not matter what we use to separate the fields, but it should be an unusual character that would not appear anywhere else in the dataset. Popular choices include the vertical bar (!) and the semicolon (;). So let's take a slight detour to find out how to convert a comma-separated file into a semicolon separated file.

You will know if you need to convert your file if, when you try to set up your SQL database, you receive an error message about an “extra column.”

We are going to use a `while` loop to read in 1,000,000 rows of the CSV file at a time. R can handle 1,000,000 rows. With 1,000,000 rows read in, we will use a regular expression to replace all the commas used for separating columns with semicolons. Then we will write out the resulting cleaned up rows into a new file. It is a big file so this code can take a few minutes to run to completion.

```
infile <- file("Crimes_-_2001_to_present.csv", 'r')           # 'r' for 'read'
outfile <- file("Crimes_-_2001_to_present-clean.csv", 'w') # 'w' for 'write'

# fix the Row #1 with the columns names
readLines(infile, n=1) |>
  gsub(", ", ";", x=_) |> # separate with ;
  gsub(" ", "", x=_) |> # SQL doesn't like field names with .,-,space
  writeLines(con=outfile)

cLines <- 0 # just a counter for the number of lines read

# read in 1000000 lines. keep going if more than 0 lines read
while ((length(a <- readLines(infile, n=1000000)) > 0))
{
  cLines <- cLines + length(a) # increase the line counter
  cLines |> format(big.mark=",", scientific=FALSE) |> message()
  # remove any semicolons if they are there
  a <- gsub(";", "", a)
  # use ?= to "lookahead" for paired quotes
  a <- gsub(",(?=(["\"]|\"[^"]*\")*$)", ";", a, perl=TRUE)
  # write the cleaned up data to storage
  writeLines(a, con=outfile)
}
```

1,000,000

2,000,000

3,000,000

4,000,000

5,000,000

6,000,000

7,000,000

8,000,000

8,390,646

```
close(infile)  
close(outfile)
```

Now, let's take a look at the first five lines of the new file we just created.

```
scan(what="", file="Crimes_-_2001_to_present-clean.csv", nlines=5, sep="\n")
```

You now see that semicolons separate the columns rather than commas. That previous record that had the location description “SCHOOL, PUBLIC, BUILDING” now looks like this:

[1] "10000153;HY189345;03/18/2015 12:20:00 PM;091XX S UNIVERSITY AVE;0483;BATTERY;AGG PRO.EMI

Note that the commas are still there inside the quotes. Now we will be able to tell SQLite to look for semicolons to separate the columns.

3 Setting up the database

Now that the csv file containing the data is ready, we can load it into SQLite.

```
# peek at the first few rows of the dataset
a <- read.table("Crimes_-_2001_to_present-clean.csv",
                 sep=";", nrows=5, header=TRUE)
# ask SQLite what data type it plans to use to store each column (eg number, text)
variabletypes <- dbDataType(con, a)
# make sure these features are stored as TEXT
variabletypes[c("IUCR", "FBICode", "Ward", "District", "CommunityArea")] <- "TEXT"

# just in case you already created a "crime" table, delete it
if(dbExistsTable(con, "crime")) dbRemoveTable(con, "crime")
# import the data file into the database
dbWriteTable(con, "crime",                               # create crime table
             "Crimes_-_2001_to_present-clean.csv", # from our cleaned up file
             row.names=FALSE,                      # first row has column names
             header=TRUE,                         # columns separated with ;
             field.types=variabletypes,
             sep=";")                           # columns separated with ;

# does the table exist?
dbListTables(con)
```

```
[1] "crime"
```

```
# a quick check to see if all the columns are there
dbListFields(con, "crime")
```

```
[1] "ID"                  "CaseNumber"          "Date"
[4] "Block"                "IUCR"                 "PrimaryType"
[7] "Description"         "LocationDescription" "Arrest"
[10] "Domestic"            "Beat"                 "District"
[13] "Ward"                 "CommunityArea"       "FBICode"
[16] "XCoordinate"         "YCoordinate"         "Year"
[19] "UpdatedOn"            "Latitude"            "Longitude"
[22] "Location"
```

```
# disconnect from the database to finalize
dbDisconnect(con)
```

You will know if the database has been successfully set up if you find a chicagocrime.db file that has about 2 Gb of data in it. If the file size is 0 or really small, then you may be looking in the wrong folder or the data cleaning and import did not finish.

```
# how many gigabytes?  
(file.size("chicagocrime.db")/10^9) |>  
  round(1) |>  
  format(nsmall=1, scientific=FALSE)
```

```
[1] "1.9"
```

Once you have successfully set up your database, there is no reason to run these lines of code again. You should never again need to turn commas into semicolons or run `dbWriteTable()`. Instead, every time you want to work with your database, you can simply need to reconnect to the database with:

```
con <- dbConnect(SQLite(), dbname="chicagocrime.db")
```

Note that if you are using a cloud-based backup service like iCloud, OneDrive, or Google Drive, you might need to wait until your “db” file has completely synced before you can access your database. For this reason I typically put my SQLite databases in a folder that does not get backed up. If I accidentally delete it, then I just rerun the code to rebuild the database.

4 SQL queries (SELECT, WHERE, FROM)

You have now created a database `chicagocrime.db` containing a table called `crime` that contains those 8 million crime records.

Two important clauses with an SQL query are `SELECT` and `FROM`. Unlike R, SQL queries are not case-sensitive and column names are not case-sensitive. So if we were to type “`SELECT`” as “`select`” or “`Description`” as “`dEsCrIpTiOn`”, the SQL query would do the same thing. However, the tradition is to put SQL keywords in all uppercase to make it easier to distinguish them from table and column names.

The `SELECT` clause tells SQL which columns in particular you would like to see. The `FROM` clause simply tells SQL from which table it should pull the data. In this query, we are interested in only the `ID` and `Description` columns.

```
dbGetQuery(con,  
  "SELECT ID, Description  
   FROM crime",  
  n = 10) # just the first 10 rows
```

	ID	Description
1	13311263	CHILD PORNOGRAPHY
2	13053066	MANUFACTURE / DELIVER - CRACK
3	12131221	AGGRAVATED VEHICULAR HIJACKING
4	11227634	NON-AGGRAVATED
5	13203321	TO VEHICLE
6	13204489	OVER \$500
7	11695116	UNLAWFUL ENTRY
8	12419690	SEXUAL EXPLOITATION OF A CHILD
9	12729745	ATTEMPT STRONG ARM - NO WEAPON
10	12835559	AUTOMOBILE

`dbGetQuery()` pulls the selected rows (first 10) from the selected columns (ID and Description). Sometimes it is preferable to get large datasets in smaller chunks using `dbSendQuery()` and `dbFetch()`.

```
res <- dbSendQuery(con, "
  SELECT ID,Description
  FROM crime")
# pull the first 10 lines
dbFetch(res, n = 10)
```

	ID	Description
1	13311263	CHILD PORNOGRAPHY
2	13053066	MANUFACTURE / DELIVER - CRACK
3	12131221	AGGRAVATED VEHICULAR HIJACKING
4	11227634	NON-AGGRAVATED
5	13203321	TO VEHICLE
6	13204489	OVER \$500
7	11695116	UNLAWFUL ENTRY
8	12419690	SEXUAL EXPLOITATION OF A CHILD
9	12729745	ATTEMPT STRONG ARM - NO WEAPON
10	12835559	AUTOMOBILE

```
# pull the next 10 lines
dbFetch(res, n = 10)
```

	ID	Description
1	13003649	FORCIBLE ENTRY
2	13061203	DOMESTIC BATTERY SIMPLE
3	13256787	DOMESTIC BATTERY SIMPLE

```

4 13116982                               RECKLESS HOMICIDE
5 13364090 "PROTECTED EMPLOYEE - HANDS, FISTS, FEET, NO / MINOR INJURY"
6 13376308      "AGGRAVATED P.O. - HANDS, FISTS, FEET, NO / MINOR INJURY"
7      27382                               FIRST DEGREE MURDER
8      27547                               FIRST DEGREE MURDER
9      6255892                           ARMED - HANDGUN
10     6272641                          STRONG ARM - NO WEAPON

```

```

# when finished, clear the rest of the results
dbClearResult(res)

```

`dbClearResult(res)` tells SQLite that we are all done with this query. We have displayed the first 20 rows. SQLite is standing by with another 8 million rows to show us, but `dbClearResult(res)` tells SQLite that we are no longer interested in this query and it can clear out whatever it has stored for us.

In the previous SQL query we just asked for `ID` and `Description`. Typing out all of the column names would be tiresome, so SQL lets you use a `*` to select all the columns. If we want to look at the first 10 rows but all of the columns, we would use this query:

```

dbGetQuery(con, "
SELECT *
FROM crime",
n = 3)

```

Warning: Column `XCoordinate`: mixed type, first seen values of type string,
coercing other values of type integer

Warning: Column `YCoordinate`: mixed type, first seen values of type string,
coercing other values of type integer

Warning: Column `Latitude`: mixed type, first seen values of type string,
coercing other values of type real

Warning: Column `Longitude`: mixed type, first seen values of type string,
coercing other values of type real

	ID	CaseNumber	Date	Block	IUCR
1	13311263	JG503434	07/29/2022 03:39:00 AM	023XX S TROY ST	1582
2	13053066	JG103252	01/03/2023 04:44:00 PM	039XX W WASHINGTON BLVD	2017
3	12131221	JD327000	08/10/2020 09:45:00 AM	015XX N DAMEN AVE	0326

	PrimaryType	Description	LocationDescription						
1	OFFENSE INVOLVING CHILDREN	CHILD PORNOGRAPHY	RESIDENCE						
2	NARCOTICS	MANUFACTURE / DELIVER - CRACK	SIDEWALK						
3	ROBBERY	AGGRAVATED VEHICULAR HIJACKING	STREET						
	Arrest	Domestic	Beat	District	Ward	CommunityArea	FBICode	XCoordinate	
1	true	false	1033	010	25		30	17	
2	true	false	1122	011	28		26	18	
3	true	false	1424	014	1		24	03	1162795
			YCoordinate	Year		UpdatedOn	Latitude	Longitude	
1				2022	04/18/2024	03:40:59 PM			
2				2023	01/20/2024	03:41:12 PM			
3			1909900	2020	05/17/2025	03:40:52 PM	41.908417822	-87.67740693	
					Location				
1						\r			
2						\r			
3	"(41.908417822,	-87.67740693)"				\r			

In addition to showing us the first three rows in their entirety, we get some warnings here regarding the coordinates of the crime that we will have to deal with later. The issue involves how SQL stores missing values.

Just as `SELECT` filters the columns, the `WHERE` clause filters the rows. Note the use of `AND` and `OR` in the `WHERE` clause. Here we select three columns: `ID`, `Description`, and `LocationDescription`. Also, we want only rows where

- the value in the `Beat` column is “611”
- the value in the `Arrest` column is “true”
- the value in the `IUCR` column is either “0486” or “0498”

Importantly, note the use of single (not double) quotation marks in the `WHERE` line. The reason is that if we used double quotes, then R will think that the double quote signals the end of the query.

```
a <- dbGetQuery(con, "
  SELECT ID, Description, LocationDescription
  FROM crime
  WHERE ((Beat=611) AND
         (Arrest='true')) AND
         ((IUCR='0486') OR (IUCR='0498')))"
# show the first few rows of the results
head(a, 3)
```

ID	Description	LocationDescription
----	-------------	---------------------

1 13248950 DOMESTIC BATTERY SIMPLE	APARTMENT
2 13254239 DOMESTIC BATTERY SIMPLE	SIDEWALK
3 13287327 DOMESTIC BATTERY SIMPLE	APARTMENT

SQLite allows regular expressions in the WHERE clause. First you have to initialize the regular expression SQL extension. Then you can insert a regular expression after the keyword REGEXP.

```
# once per R session initialize regexp
initExtension(con, "regexp")
# get crimes from beats that start with "12"
a <- dbGetQuery(con, "
  SELECT Beat
  FROM   crime
  WHERE  Beat REGEXP '^12'..$',
  n = -1)

unique(a$Beat)
```

```
[1] 122 123 224 232 133 222 132 215 124 211 221 114 225 214 131 231 112 113 233
[20] 111 234 235 121 213 223 212 134
```

There is a full list of all available [SQLite extensions](#). Frankly, I have only ever used the REGEXP extension.

SQL does not like column names with special characters. Only letters (first character *must* be a letter), numbers, and underscores (_). Column names also cannot be a SQL keyword, like SELECT or WHERE. If you happen to have a table with any special characters, like periods, hyphens, or spaces, you can “protect” the column name in square brackets. For example, SELECT [incident id], [text-description], [location.description], [where].

4.1 Exercises

1. Select records from Beat 234
2. Select Beat, District, Ward, and Community Area for all “ASSAULT”s
3. Select records on assaults from Beat 234
4. Make a table of the number of assaults (IUCR 0560) by Ward

5 GROUP BY and aggregation functions

We have already covered SQL clauses SELECT, WHERE, and FROM. The SQL function COUNT(*) and GROUP BY are also very useful. For example, the following query counts how many assaults (IUCR 0560) occurred by ward. COUNT() is a SQL “aggregate” function, a function that performs a calculation on a group of values and returns a single number. Other SQL aggregate functions include AVG(), MIN(), MAX(), and SUM(). This query will group all the records by Ward and then apply the aggregate function COUNT() and report that value in a column called crimecount. AS allows us to give clear column names in the results. Without the AS crimecount column of counts would be called COUNT(*), which has several characters about which SQL will complain.

```
a <- dbGetQuery(con, "
  SELECT COUNT(*) AS crimecount,
         Ward
  FROM crime
  WHERE IUCR='0560'
  GROUP BY Ward")
print(a)
```

	crimecount	Ward
1	29470	
2	5306	1
3	8100	10
4	5085	11
5	4273	12
6	3859	13
7	4265	14
8	9556	15
9	11339	16
10	13662	17
11	6130	18
12	3546	19
13	10779	2
14	13107	20
15	11488	21
16	4350	22
17	4083	23
18	12604	24
19	5326	25
20	6479	26
21	11721	27

22	15148	28
23	8899	29
24	11767	3
25	4448	30
26	4499	31
27	3244	32
28	3004	33
29	11107	34
30	4468	35
31	3738	36
32	9354	37
33	3424	38
34	2935	39
35	8565	4
36	3734	40
37	3233	41
38	9691	42
39	2191	43
40	3173	44
41	3617	45
42	5200	46
43	2695	47
44	3967	48
45	5195	49
46	9269	5
47	3326	50
48	12936	6
49	11797	7
50	11727	8
51	11744	9

The GROUP BY clause is critical. If you forget it then the result is not well defined. That is, different implementations of SQL may produce different results. The rule you should remember is that “every non-aggregated column in the SELECT clause should appear in the GROUP BY clause.” Here *Ward* is not part of the aggregate function COUNT() so it must appear in the GROUP BY clause.

5.1 Exercises

5. Count the number of crimes by PrimaryType
6. Count the number of crimes resulting in arrest

- Count the number of crimes by `LocationDescription`. `LocationDescription` is the variable that tells us where (e.g., a parking lot, a barbershop, a fire station, a CTA train, or a motel) a crime occurred

6 ORDER BY and UPDATE

`MAX`, `MIN`, `SUM`, `AVG` are common (and useful) aggregating functions. The `ORDER BY` clause sorts the results for us. It is the SQL version of the `sort()` or `arrange()` functions. Here is an illustration that gives the range of beat numbers in each policing district.

```
dbGetQuery(con, "
  SELECT MIN(Beat) AS min_beat,
         MAX(Beat) AS max_beat,
         District
    FROM crime
   GROUP BY District
  ORDER BY District")
```

	min_beat	max_beat	District
1	124	2535	
2	111	2535	001
3	131	2232	002
4	133	2222	003
5	324	2514	004
6	333	2233	005
7	123	2424	006
8	233	2431	007
9	333	2411	008
10	131	2522	009
11	133	2534	010
12	624	2535	011
13	111	2525	012
14	411	2535	014
15	726	2533	015
16	811	2521	016
17	734	2523	017
18	111	2533	018
19	112	2533	019
20	112	2433	020
21	2112	2112	021
22	214	2234	022

```

23      123      2433      024
24      725      2535      025
25      124      2535      031
26      1614     1614      16

```

Remember that the `GROUP BY` clause should include every element of the `SELECT` clause that is not involved with an aggregate function. We have `MIN()` and `MAX()` operating on `Beat`, but `District` is on its own and should be placed in the `GROUP BY` clause.

Let's look at our `Latitude` and `Longitude` columns, which will be extremely useful for mapping data points. The following query will give unexpected results.

```

dbGetQuery(con, "
  SELECT MIN(Latitude) AS min_lat,
         MAX(Latitude) AS max_lat,
         MIN(Longitude) AS min_lon,
         MAX(Longitude) AS max_lon,
         District
    FROM crime
   GROUP BY District
  ORDER BY District")

```

Warning: Column `max_lat`: mixed type, first seen values of type real, coercing other values of type string

Warning: Column `max_lon`: mixed type, first seen values of type real, coercing other values of type string

	min_lat	max_lat	min_lon	max_lon	District
1	41.69991	42.00030	-87.87742	-87.59533	
2	36.61945	0.00000	-91.68657	0.00000	001
3	36.61945	0.00000	-91.68657	0.00000	002
4	36.61945	0.00000	-91.68657	0.00000	003
5	36.61945	0.00000	-91.68657	0.00000	004
6	36.61945	0.00000	-91.68657	0.00000	005
7	36.61945	0.00000	-91.68657	0.00000	006
8	36.61945	0.00000	-91.68657	0.00000	007
9	36.61945	0.00000	-91.68657	0.00000	008
10	36.61945	0.00000	-91.68657	0.00000	009
11	36.61945	0.00000	-91.68657	0.00000	010
12	36.61945	0.00000	-91.68657	0.00000	011
13	36.61945	0.00000	-91.68657	0.00000	012

```

14 36.61945 0.00000 -91.68657 0.00000 014
15 36.61945 0.00000 -91.68657 0.00000 015
16 36.61945 0.00000 -91.68657 0.00000 016
17 36.61945 0.00000 -91.68657 0.00000 017
18 36.61945 0.00000 -91.68657 0.00000 018
19 41.80933 0.00000 -87.76791 0.00000 019
20 41.79145 0.00000 -87.76303 0.00000 020
21 41.83790 41.83790 -87.62192 -87.62192 021
22 36.61945 0.00000 -91.68657 0.00000 022
23 36.61945 0.00000 -91.68657 0.00000 024
24 36.61945 0.00000 -91.68657 0.00000 025
25 41.64619 42.01939 -87.93973 -87.53528 031
26 41.98531 41.98552 -87.83047 -87.82900 16

```

We get some strange results here. `max_lat` equal to 0.0 is on the equator! It is doubtful that Chicago reported any equatorial crimes. The problem is that we have some blank values in `Longitude` and `Latitude`. Here are some of them.

```
dbGetQuery(con, "SELECT * FROM crime WHERE Longitude=''", n=3)
```

ID	CaseNumber	Date	Block	IUCR			
1	JG503434	07/29/2022 03:39:00 AM	023XX S TROY ST	1582			
2	JG103252	01/03/2023 04:44:00 PM	039XX W WASHINGTON BLVD	2017			
3	JB147599	08/26/2017 10:00:00 AM	001XX W RANDOLPH ST	0281			
	PrimaryType	Description	LocationDescription				
1	OFFENSE INVOLVING CHILDREN	CHILD PORNOGRAPHY	RESIDENCE				
2	NARCOTICS MANUFACTURE / DELIVER - CRACK		SIDEWALK				
3	CRIM SEXUAL ASSAULT	NON-AGGRAVATED	HOTEL/MOTEL				
Arrest	Domestic	Beat	District	Ward	CommunityArea	FBICode	XCoordinate
1	true	false	1033	010	25	30	17
2	true	false	1122	011	28	26	18
3	false	false	122	001	42	32	02
YCoordinate	Year	UpdatedOn	Latitude	Longitude	Location		
1		2022 04/18/2024 03:40:59 PM			\r		
2		2023 01/20/2024 03:41:12 PM			\r		
3		2017 02/11/2018 03:57:41 PM			\r		

Note that the `Latitude` and the `Longitude` columns are blank. Also, look at these

```
dbGetQuery(con, "SELECT * FROM crime where Latitude<36.61946", n=3)
```

ID	CaseNumber	Date	Block	IUCR	PrimaryType
1	1482	HH367441 05/13/2002 05:00:00 AM	061XX S ARTESIAN ST	0110	HOMICIDE
2	838	G311269 05/29/2001 11:35:00 PM	059XX S MORGAN AV	0110	HOMICIDE
3	637	G005960 01/06/2001 10:35:00 AM	014XX N HARDING ST	0110	HOMICIDE
Description LocationDescription Arrest Domestic Beat District Ward					
1	FIRST DEGREE MURDER	HOUSE	true	false	825 008
2	FIRST DEGREE MURDER	DUMPSTER	true	false	712 007
3	FIRST DEGREE MURDER	STREET	true	false	2535 025
CommunityArea	FBICode	XCoordinate	YCoordinate	Year	UpdatedOn
1	01A	0	0	2002	01/28/2024 03:40:59 PM
2	01A	0	0	2001	01/28/2024 03:40:59 PM
3	01A	0	0	2001	01/28/2024 03:40:59 PM
Latitude	Longitude	Location			
1	36.61945	-91.68657	"(36.619446395, -91.686565684)"\r		
2	36.61945	-91.68657	"(36.619446395, -91.686565684)"\r		
3	36.61945	-91.68657	"(36.619446395, -91.686565684)"\r		

The point (-91.68657, 36.61945) lands in Brandsville, Missouri, also a highly unlikely location for Chicago crime.

We can tell SQLite to make the empty or missing values NULL, a more proper way to encode that these rows have missing coordinates. The UPDATE clause edits our table. R will read in NULL values as NA. After we do the update, we can rerun the MIN(), MAX() query. We can also assign NULL to latitudes and longitudes that are very close to 0.

Note that we use dbExecute() when updating since we are not asking for any rows of data to come back to us.

```
dbExecute(con, "
  UPDATE crime SET Latitude=NULL
  WHERE (Latitude='') OR (ABS(Latitude-0.0) < 0.01) OR (Latitude < 36.7)")
```

[1] 93655

```
dbExecute(con, "
  UPDATE crime SET Longitude=NULL
  WHERE (Longitude='') OR (ABS(Longitude-0.0) < 0.01) OR (Longitude < -91.6)")
```

[1] 93655

Let's rerun that query and check that we get more sensible results.

```

dbGetQuery(con, "
  SELECT MIN(Latitude) AS min_lat,
         MAX(Latitude) AS max_lat,
         MIN(Longitude) AS min_lon,
         MAX(Longitude) AS max_lon,
         District
    FROM crime
   GROUP BY District
  ORDER BY District")

```

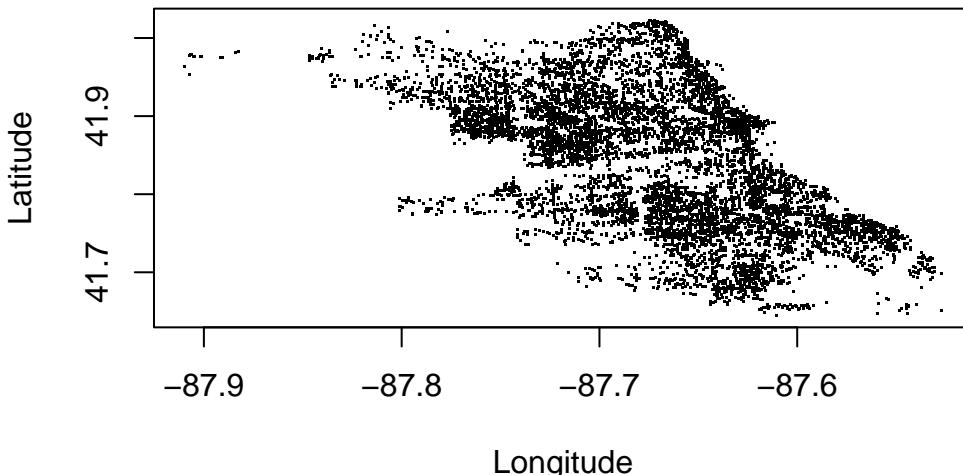
	min_lat	max_lat	min_lon	max_lon	District
1	41.69991	42.00030	-87.87742	-87.59533	
2	41.72827	41.98740	-87.84349	-87.54925	001
3	41.73298	41.97608	-87.70277	-87.56954	002
4	41.71424	41.79946	-87.73941	-87.55261	003
5	41.64467	41.79220	-87.72436	-87.52453	004
6	41.64459	41.88693	-87.73145	-87.54348	005
7	41.69249	42.01876	-87.77138	-87.55810	006
8	41.66806	42.01369	-87.68723	-87.57906	007
9	41.73453	42.01765	-87.80161	-87.55239	008
10	41.77015	41.97645	-87.71397	-87.60282	009
11	41.68357	41.94304	-87.74364	-87.61895	010
12	41.77163	41.90624	-87.76332	-87.62328	011
13	41.68544	41.96539	-87.76321	-87.60502	012
14	41.77688	42.01938	-87.80222	-87.65657	014
15	41.76641	41.94234	-87.77535	-87.63087	015
16	41.78464	42.01938	-87.93457	-87.58256	016
17	41.77950	42.01390	-87.75780	-87.66131	017
18	41.85952	41.96879	-87.76313	-87.60136	018
19	41.80933	41.98397	-87.76791	-87.58775	019
20	41.79145	42.00458	-87.76303	-87.62992	020
21	41.83790	41.83790	-87.62192	-87.62192	021
22	41.67709	41.85572	-87.74328	-87.58965	022
23	41.75988	42.02291	-87.79757	-87.62545	024
24	41.83930	41.94586	-87.81648	-87.64093	025
25	41.64619	42.01939	-87.93973	-87.53528	031
26	41.98531	41.98552	-87.83047	-87.82900	16

Now we have results that are more in line with where Chicago actually is. Make it a habit to do some checks of your data before doing too much analysis.

And what city does the following plot have the shape of? Let's plot the location of these crimes. Plotting all 8 million would be overkill, so let's take a random sample of 10,000 crimes. Here is a SQL query that will randomly order the rows and select just the first 10,000. Does the shape of the plot look right?

```
a <- dbGetQuery(con, "
  SELECT Longitude, Latitude
  FROM crime
  ORDER BY RANDOM() -- scramble the order of the rows
  LIMIT 10000")

plot(Latitude~Longitude, data=a,
  pch=".",
  xlab="Longitude", ylab="Latitude")
```



6.1 Exercises

8. Plot the longitude and latitude of all “ASSAULT”s for Ward 22
9. What is the most common (Long,Lat) for assaults in Ward 22? Add the point to your plot using the `points()` function. `points()` simply draws a point (or sequence of points) at the specified coordinates

And always disconnect when you are done.

```
dbDisconnect(con)
```

7 Solutions to the exercises

1. Select records from Beat 234

```
dbGetQuery(con, "
SELECT *
FROM crime
WHERE Beat=234",
n=5)
```

Warning: Column `XCoordinate`: mixed type, first seen values of type integer, coercing other values of type string

Warning: Column `YCoordinate`: mixed type, first seen values of type integer, coercing other values of type string

ID	CaseNumber	Date	Block	IUCR				
1	13208531	JG408244 08/01/2023 12:00:00 PM	054XX S EAST VIEW PARK	0820				
2	13203370	JG415497 09/07/2023 07:30:00 PM	051XX S KENWOOD AVE	1310				
3	13207450	JG420345 09/07/2023 01:54:00 PM	054XX S BLACKSTONE AVE	0890				
4	13203210	JG415469 09/07/2023 06:30:00 PM	052XX S BLACKSTONE AVE	0890				
5	13206379	JG418537 01/01/2007 04:40:00 PM	053XX S SHORE DR	1153				
	PrimaryType		Description	LocationDescription				
1	THEFT	\$500 AND UNDER		STREET				
2	CRIMINAL DAMAGE	TO PROPERTY		APARTMENT				
3	THEFT	FROM BUILDING		APARTMENT				
4	THEFT	FROM BUILDING		APARTMENT				
5	DECEPTIVE PRACTICE FINANCIAL IDENTITY THEFT OVER \$ 300							
Arrest	Domestic	Beat	District	Ward	CommunityArea	FBICode	XCoordinate	
1	false	false	234	002	5	41	06	1188934
2	false	true	234	002	4	41	14	1185980
3	false	false	234	002	5	41	06	1186841
4	false	false	234	002	4	41	06	1186800
5	false	false	234	002	5	41	11	0
YCoordinate	Year		UpdatedOn	Latitude	Longitude			
1	1869643	2023	09/14/2023 03:41:59 PM	41.79736	-87.58268			

```

2 1871242 2023 09/15/2023 03:42:23 PM 41.80182 -87.59346
3 1869253 2023 09/15/2023 03:42:23 PM 41.79634 -87.59037
4 1870814 2023 09/15/2023 03:42:23 PM 41.80063 -87.59047
5 0 2007 09/16/2023 03:42:58 PM NA NA
      Location
1 "(41.79736226, -87.582679493)"\r
2 "(41.801820311, -87.593461583)"\r
3 "(41.796341968, -87.590367054)"\r
4 "(41.80062644, -87.590467932)"\r
5 \r

```

2. Select Beat, District, Ward, and Community Area for all “ASSAULT”s

```

dbGetQuery(con, "
  SELECT Beat, District, Ward, CommunityArea, PrimaryType
  FROM crime
  WHERE PrimaryType='ASSAULT'",
  n=5)

```

	Beat	District	Ward	CommunityArea	PrimaryType
1	2515	025	36	19	ASSAULT
2	1713	017	33	14	ASSAULT
3	631	006	6	44	ASSAULT
4	322	003	6	69	ASSAULT
5	1533	015	29	25	ASSAULT

3. Select records on assaults from Beat 234

```

dbGetQuery(con, "
  SELECT *
  FROM crime
  WHERE (Beat=234) AND (PrimaryType='ASSAULT')",
  n=5)

```

	ID	CaseNumber	Date	Block	IUCR		
1	13276965	JG502615	11/10/2023 09:00:00 AM	015XX E HYDE PARK BLVD	0560		
2	13207370	JG420456	09/10/2023 05:19:00 PM	053XX S HYDE PARK BLVD	0560		
3	13210166	JG421339	09/12/2023 01:30:00 PM	015XX E 53RD ST	0560		
4	13273166	JG499223	11/10/2023 04:39:00 PM	015XX E 53RD ST	0560		
5	13225905	JG442370	09/11/2023 04:15:00 PM	054XX S CORNELL AVE	0560		
	PrimaryType	Description	LocationDescription	Arrest	Domestic	Beat	District
1	ASSAULT	SIMPLE	ATHLETIC CLUB	false	false	234	002

```

2 ASSAULT SIMPLE APARTMENT false false 234 002
3 ASSAULT SIMPLE STREET false false 234 002
4 ASSAULT SIMPLE SMALL RETAIL STORE false false 234 002
5 ASSAULT SIMPLE APARTMENT false false 234 002
Ward CommunityArea FBICode XCoordinate YCoordinate Year
1 4 41 08A 1187293 1871488 2023
2 5 41 08A 1188556 1870311 2023
3 4 41 08A 1187634 1870434 2023
4 5 41 08A 1187748 1870436 2023
5 5 41 08A 1188178 1869513 2023
UpdatedOn Latitude Longitude Location
1 11/18/2023 03:40:25 PM 41.80246 -87.58864 "(41.802464238, -87.588638554)"\r
2 09/18/2023 03:42:32 PM 41.79920 -87.58404 "(41.799204348, -87.584044296)"\r
3 09/20/2023 03:42:29 PM 41.79956 -87.58742 "(41.799563873, -87.587421525)"\r
4 11/18/2023 03:40:25 PM 41.79957 -87.58700 "(41.799566646, -87.5870034)"\r
5 09/30/2023 03:41:20 PM 41.79702 -87.58546 "(41.797023613, -87.585455951)"\r

```

4. Make a table of the number of assaults (IUCR 0560) by Ward

We could select all the IUCR codes and ward with SQL and then filter and tabulate the data in R.

```

# system.time() reports how long it takes to run the SQL query
# How long if we retrieve data from SQL and tabulate in R?
system.time(
{
  data <- dbGetQuery(con, "
    SELECT IUCR,Ward
    FROM crime")
  data |>
    filter(IUCR=="0560") |>
    count(Ward)
})

```

```

user  system elapsed
4.03    1.70    5.75

```

Or we could make SQL do all the work selecting and tabulating.

```

# How long if we make SQL do all the work?
system.time(
{

```

```

a <- dbGetQuery(con, "
  SELECT COUNT(*) AS crimecount,
         Ward
    FROM crime
   WHERE IUCR='0560'
  GROUP BY Ward")
})

```

```

user  system elapsed
0.84    2.03   2.91

```

Generally, SQL will be much faster for general selecting, filtering, tabulating, and linking data.

5. Count the number of crimes by PrimaryType

```

dbGetQuery(con, "
  SELECT COUNT(*) AS crimecount,
         PrimaryType
    FROM crime
  GROUP BY PrimaryType")

```

	crimecount	PrimaryType
1	14368	ARSON
2	561234	ASSAULT
3	1528740	BATTERY
4	443840	BURGLARY
5	1620	CONCEALED CARRY LICENSE VIOLATION
6	27296	CRIM SEXUAL ASSAULT
7	954180	CRIMINAL DAMAGE
8	11220	CRIMINAL SEXUAL ASSAULT
9	225885	CRIMINAL TRESPASS
10	385782	DECEPTIVE PRACTICE
11	1	DOMESTIC VIOLENCE
12	14661	GAMBLING
13	13906	HOMICIDE
14	139	HUMAN TRAFFICKING
15	20090	INTERFERENCE WITH PUBLIC OFFICER
16	5077	INTIMIDATION
17	7483	KIDNAPPING
18	15357	LIQUOR LAW VIOLATION

19	428309	MOTOR VEHICLE THEFT
20	762793	NARCOTICS
21	38	NON - CRIMINAL
22	190	NON-CRIMINAL
23	9	NON-CRIMINAL (SUBJECT SPECIFIED)
24	945	OBSCENITY
25	60232	OFFENSE INVOLVING CHILDREN
26	162	OTHER NARCOTIC VIOLATION
27	522986	OTHER OFFENSE
28	70379	PROSTITUTION
29	215	PUBLIC INDECENCY
30	54633	PUBLIC PEACE VIOLATION
31	24	RITUALISM
32	313817	ROBBERY
33	34026	SEX OFFENSE
34	6052	STALKING
35	1780782	THEFT
36	124175	WEAPONS VIOLATION

6. Count the number of crimes resulting in arrest

```
dbGetQuery(con, "
SELECT COUNT(*) AS crimecount, PrimaryType
FROM crime
WHERE Arrest='true'
GROUP BY PrimaryType")
```

	crimecount	PrimaryType
1	1774	ARSON
2	113725	ASSAULT
3	331346	BATTERY
4	25365	BURGLARY
5	1565	CONCEALED CARRY LICENSE VIOLATION
6	4365	CRIM SEXUAL ASSAULT
7	62083	CRIMINAL DAMAGE
8	816	CRIMINAL SEXUAL ASSAULT
9	153649	CRIMINAL TRESPASS
10	47750	DECEPTIVE PRACTICE
11	1	DOMESTIC VIOLENCE
12	14555	GAMBLING
13	6671	HOMICIDE
14	13	HUMAN TRAFFICKING

```

15      18407    INTERFERENCE WITH PUBLIC OFFICER
16          731        INTIMIDATION
17          798        KIDNAPPING
18      15207        LIQUOR LAW VIOLATION
19      32533        MOTOR VEHICLE THEFT
20      757836       NARCOTICS
21          6        NON - CRIMINAL
22          18       NON-CRIMINAL
23          3   NON-CRIMINAL (SUBJECT SPECIFIED)
24          700      OBSCENITY
25      11643      OFFENSE INVOLVING CHILDREN
26          108     OTHER NARCOTIC VIOLATION
27      92385      OTHER OFFENSE
28      70068      PROSTITUTION
29          211     PUBLIC INDECENCY
30      34107      PUBLIC PEACE VIOLATION
31          3        RITUALISM
32      29017      ROBBERY
33          8666    SEX OFFENSE
34          731      STALKING
35      193124     THEFT
36      90159      WEAPONS VIOLATION

```

Or, if we were not interested in differentiating based on the PrimaryType, we could simply do the following:

```

dbGetQuery(con, "
  SELECT COUNT(*) AS crimecount
  FROM crime
  WHERE Arrest='true'")

```

```

  crimecount
1      2120139

```

7. Count the number of crimes by LocationDescription

```

dbGetQuery(con, "
  SELECT COUNT(*) AS crimecount, LocationDescription
  FROM crime
  GROUP BY LocationDescription
  ORDER BY crimecount DESC")

```

	crimecount	LocationDescription
1	2192032	STREET
2	1379598	RESIDENCE
3	994878	APARTMENT
4	760617	SIDEWALK
5	269956	OTHER
6	202933	PARKING LOT/GARAGE(NON.RESID.)
7	186884	ALLEY
8	168253	SMALL RETAIL STORE
9	146368	"SCHOOL, PUBLIC, BUILDING"
10	140867	RESTAURANT
11	135279	RESIDENCE-GARAGE
12	132886	VEHICLE NON-COMMERCIAL
13	124166	RESIDENCE PORCH/HALLWAY
14	110961	DEPARTMENT STORE
15	104800	GROCERY FOOD STORE
16	93644	GAS STATION
17	75140	RESIDENTIAL YARD (FRONT/BACK)
18	69131	COMMERCIAL / BUSINESS OFFICE
19	63309	PARK PROPERTY
20	56098	CHA PARKING LOT/GROUNDS
21	47020	BAR OR TAVERN
22	44857	PARKING LOT / GARAGE (NON RESIDENTIAL)
23	41261	CTA PLATFORM
24	40430	CHA APARTMENT
25	39609	DRUG STORE
26	34389	CTA TRAIN
27	33058	BANK
28	30249	"SCHOOL, PUBLIC, GROUNDS"
29	29723	HOTEL/MOTEL
30	27692	CONVENIENCE STORE
31	27417	CTA BUS
32	25021	CHA HALLWAY/STAIRWELL/ELEVATOR
33	24679	VACANT LOT/LAND
34	24674	DRIVEWAY - RESIDENTIAL
35	23050	OTHER (SPECIFY)
36	22458	TAVERN/LIQUOR STORE
37	22189	HOSPITAL BUILDING/GROUNDS
38	18560	POLICE FACILITY/VEH PARKING LOT
39	17707	RESIDENCE - PORCH / HALLWAY
40	16296	AIRPORT/AIRCRAFT
41	15488	CHURCH/SYNAGOGUE/PLACE OF WORSHIP
42	15296	RESIDENCE - YARD (FRONT / BACK)

43	14770	
44	14757	GOVERNMENT BUILDING/PROPERTY
45	14656	NURSING HOME/RETIREMENT HOME
46	14460	RESIDENCE - GARAGE
47	14329	CONSTRUCTION SITE
48	14182	"SCHOOL, PRIVATE, BUILDING"
49	12409	CURRENCY EXCHANGE
50	12182	ABANDONED BUILDING
51	10698	WAREHOUSE
52	10276	CTA GARAGE / OTHER PROPERTY
53	10258	ATHLETIC CLUB
54	8824	CTA BUS STOP
55	8779	BARBERSHOP
56	8707	ATM (AUTOMATIC TELLER MACHINE)
57	7838	CTA STATION
58	7816	TAXICAB
59	7582	SCHOOL - PUBLIC BUILDING
60	7487	HOSPITAL BUILDING / GROUNDS
61	7477	LIBRARY
62	7429	MEDICAL/DENTAL OFFICE
63	6892	FACTORY/MANUFACTURING BUILDING
64	6753	SCHOOL - PUBLIC GROUNDS
65	6727	HOTEL / MOTEL
66	5930	OTHER RAILROAD PROP / TRAIN DEPOT
67	5787	COLLEGE/UNIVERSITY GROUNDS
68	5644	AIRPORT TERMINAL UPPER LEVEL - SECURE AREA
69	5615	VEHICLE-COMMERCIAL
70	5335	CLEANING STORE
71	5290	SPORTS ARENA/STADIUM
72	4292	"SCHOOL, PRIVATE, GROUNDS"
73	4243	POLICE FACILITY / VEHICLE PARKING LOT
74	4090	NURSING / RETIREMENT HOME
75	3877	VACANT LOT / LAND
76	3764	DAY CARE CENTER
77	3603	CAR WASH
78	3576	OTHER COMMERCIAL TRANSPORTATION
79	2822	TAVERN / LIQUOR STORE
80	2732	MOVIE HOUSE/THEATER
81	2703	GOVERNMENT BUILDING / PROPERTY
82	2676	AIRPORT TERMINAL LOWER LEVEL - NON-SECURE AREA
83	2603	APPLIANCE STORE
84	2377	CHA PARKING LOT / GROUNDS
85	2369	CHURCH / SYNAGOGUE / PLACE OF WORSHIP

86	1876	AIRPORT PARKING LOT
87	1668	MEDICAL / DENTAL OFFICE
88	1641	AUTO / BOAT / RV DEALERSHIP
89	1532	AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA
90	1505	SCHOOL - PRIVATE GROUNDS
91	1399	COLLEGE/UNIVERSITY RESIDENCE HALL
92	1389	AUTO
93	1321	AIRPORT TERMINAL UPPER LEVEL - NON-SECURE AREA
94	1313	FIRE STATION
95	1306	JAIL / LOCK-UP FACILITY
96	1306	AIRPORT EXTERIOR - NON-SECURE AREA
97	1286	VEHICLE - COMMERCIAL
98	1181	LAKEFRONT/WATERFRONT/RIVERBANK
99	1169	COIN OPERATED MACHINE
100	1155	AIRPORT TERMINAL LOWER LEVEL - SECURE AREA
101	1141	SCHOOL - PRIVATE BUILDING
102	1084	HIGHWAY/EXPRESSWAY
103	1080	FEDERAL BUILDING
104	1012	AIRPORT VENDING ESTABLISHMENT
105	1005	POOL ROOM
106	981	AIRCRAFT
107	962	DELIVERY TRUCK
108	924	AIRPORT BUILDING NON-TERMINAL - SECURE AREA
109	921	CTA PARKING LOT / GARAGE / OTHER PROPERTY
110	910	ANIMAL HOSPITAL
111	898	CHA HALLWAY / STAIRWELL / ELEVATOR
112	825	BOWLING ALLEY
113	763	PAWN SHOP
114	760	SPORTS ARENA / STADIUM
115	745	OTHER RAILROAD PROPERTY / TRAIN DEPOT
116	722	FACTORY / MANUFACTURING BUILDING
117	706	HOUSE
118	698	BOAT/WATERCRAFT
119	597	AIRPORT EXTERIOR - SECURE AREA
120	596	"VEHICLE - OTHER RIDE SHARE SERVICE (LYFT, UBER, ETC.)"
121	592	CREDIT UNION
122	526	LAKEFRONT / WATERFRONT / RIVERBANK
123	520	BRIDGE
124	472	FOREST PRESERVE
125	467	"VEHICLE - OTHER RIDE SHARE SERVICE (E.G., UBER, LYFT)"
126	437	CEMETARY
127	430	VEHICLE - DELIVERY TRUCK
128	407	PORCH

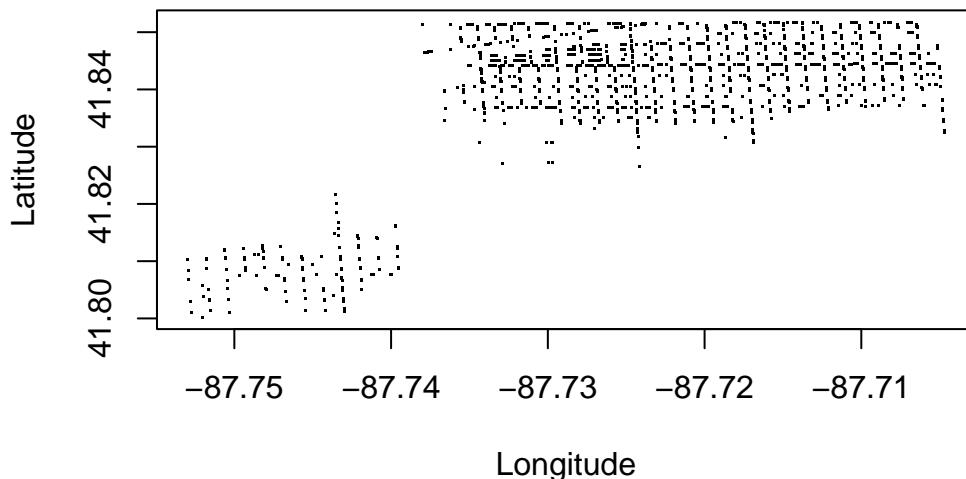
129	404	COLLEGE / UNIVERSITY - GROUNDS
130	396	SAVINGS AND LOAN
131	383	MOVIE HOUSE / THEATER
132	331	VEHICLE - OTHER RIDE SERVICE
133	330	YARD
134	286	PARKING LOT
135	272	HIGHWAY / EXPRESSWAY
136	245	NEWSSTAND
137	202	CTA TRACKS - RIGHT OF WAY
138	175	AIRPORT TRANSPORTATION SYSTEM (ATS)
139	167	BOAT / WATERCRAFT
140	152	AIRPORT TERMINAL MEZZANINE - NON-SECURE AREA
141	144	VACANT LOT
142	123	COLLEGE / UNIVERSITY - RESIDENCE HALL
143	111	HALLWAY
144	103	RETAIL STORE
145	84	CASINO/GAMBLING ESTABLISHMENT
146	75	GARAGE
147	75	GANGWAY
148	71	GAS STATION DRIVE/PROP.
149	60	CHA PARKING LOT
150	51	CHA GROUNDS
151	40	TAVERN
152	39	CHA HALLWAY
153	35	BASEMENT
154	29	DRIVEWAY
155	28	VESTIBULE
156	27	STAIRWELL
157	27	HOTEL
158	27	BARBER SHOP/BEAUTY SALON
159	22	OFFICE
160	20	VEHICLE - COMMERCIAL: TROLLEY BUS
161	20	KENNEL
162	19	HOSPITAL
163	18	RAILROAD PROPERTY
164	18	CLUB
165	17	VEHICLE - COMMERCIAL: ENTERTAINMENT / PARTY BUS
166	17	SCHOOL YARD
167	13	LIQUOR STORE
168	13	"CTA ""L"" PLATFORM"
169	11	GARAGE/AUTO REPAIR
170	11	FARM
171	11	CTA PROPERTY

172	11	"CTA ""L"" TRAIN"
173	10	VEHICLE-COMMERCIAL - TROLLEY BUS
174	10	VEHICLE-COMMERCIAL - ENTERTAINMENT/PARTY BUS
175	10	CHA STAIRWELL
176	9	TRUCK
177	9	CHA LOBBY
178	7	WOODED AREA
179	7	MOTEL
180	7	DUMPSTER
181	6	TAXI CAB
182	6	RIVER BANK
183	6	NURSING HOME
184	6	CHURCH
185	5	LAKE
186	4	TRAILER
187	4	RIVER
188	4	CHA PLAY LOT
189	3	YMCA
190	3	SEWER
191	3	HORSE STABLE
192	3	COACH HOUSE
193	3	CHA ELEVATOR
194	3	CHA BREEZEWAY
195	2	ROOMING HOUSE
196	2	PUBLIC HIGH SCHOOL
197	2	PUBLIC GRAMMAR SCHOOL
198	2	PRAIRIE
199	2	LIVERY STAND OFFICE
200	2	LAUNDRY ROOM
201	2	GOVERNMENT BUILDING
202	2	FACTORY
203	2	ELEVATOR
204	2	CTA SUBWAY STATION
205	2	COUNTY JAIL
206	2	CHURCH PROPERTY
207	2	BANQUET HALL
208	1	TRUCKING TERMINAL
209	1	ROOF
210	1	POOLROOM
211	1	POLICE FACILITY
212	1	LOADING DOCK
213	1	LIVERY AUTO
214	1	LAGOON

215	1	JUNK YARD/GARBAGE DUMP
216	1	FUNERAL PARLOR
217	1	EXPRESSWAY EMBANKMENT
218	1	CLEANERS/LAUNDROMAT
219	1	BEACH

8. Plot the longitude and latitude of all “ASSAULT”s for Ward 22

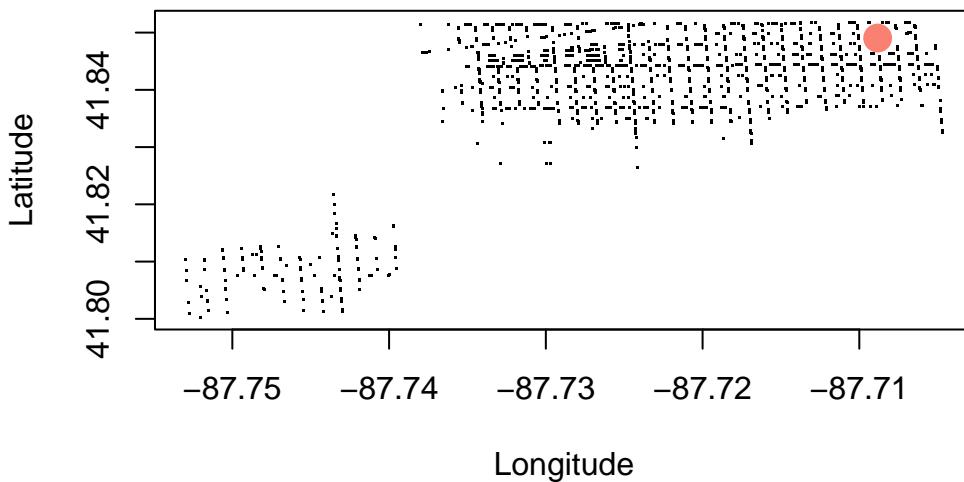
```
a <- dbGetQuery(con, "
  SELECT Latitude, Longitude
  FROM crime
  WHERE PrimaryType='ASSAULT' AND Ward='22'")
plot(Latitude~Longitude, data=a, pch=".")
```



9. What is the most common (Long,Lat) for assaults in Ward 22?

```
b <- dbGetQuery(con, "
  SELECT COUNT(*) AS crimecount,
         Latitude, Longitude
  FROM   crime
  WHERE  PrimaryType='ASSAULT' AND Ward=22
  GROUP BY Latitude, Longitude
  ORDER BY crimecount DESC")
```

```
LIMIT 1")  
  
plot(Latitude~Longitude, data=a, pch=".")  
points(Latitude~Longitude,  
       data=b,  
       pch=16,  
       col="salmon",  
       cex=2)
```



b

```
crimecount Latitude Longitude  
1          229 41.84905 -87.70883
```

```
[1] TRUE
```