# Data Storage and Retrieval – Assignment 1

*Aaron Niskin*

*August 25, 2016*

## Question 1: The directory /usr/share/databases/Homelessness/Jail/LayoutBFiles contains

- the raw data files "N LAYOUT B CSV.csv", N=1,. . . ,9
- several intermediate files of processed data, some of which are explained in the README
- some *.clean files that are pasted together into the final files bookings.clean and booking_addl-charge.clean

a. Write a program to read in the raw data files and produce the two files release.clean and booking_addl-charge.clean. Test your output with the Linux diff command.

So my `release.clean` file can come out without any differences, or it can not include lines with missing values (like release dates, etc.). This can be done just by switching a + to a * in the regular expression. It's pretty cool. But my `bookings_addl-charges.clean` file is another story. I should have just written the whole thing in python. But I realized that a bit too late. I started learning sed, awk, and perl, and that was where my weekend went.

b. Write a paragraph discussing the choice I made to separate the information in bookings.clean from the information in booking_addl-charge.clean. Was it a convenient choice for cleaning the data? for asking questions of the data?

That's called data-base normalization. Basically, you want a separation of concerns. If one thing is repeated possibly many times for every one occurrance of another thing, you'd want them both separated into different tables. My only question is why you didn't choose to include the booking ID in the `release.clean` file. It seems as though that would make a reliable ID to join on, etc.

## Question 2: Last week we asked the following questions about the jail data. For each question, write one sql statement or a sequence of sql statements that come as close to answering the question as you can. Show the top ten rows of results for each question, where you get to pick what "top" means.

Some of the questions are worded precisely; others not. If a question isn't worded precisely, give it a precise interpretation before answering it. You may embellish each question as you see fit; just say what question you are answering.

Along with your sql and results, write a paragraph about each question describing how fully your sql answers the question and whether and how additional data might help you answer the question more fully.

a. How many arrests are there for each race and ethnicity?

Let me start by apologizing for the lack of code-blocks and syntax highlighting. I haven't yet figured out how to connect SQLite to RStudio (which is apparently necessary, even just to include the code chunks).

```sql
SELECT UPPER(e), COUNT(bookingNumber) FROM bookingsB
GROUP BY UPPER(e);
```

```
H|190627

N|1031072

U|1092
```

---

```sql
SELECT race, COUNT(bookingNumber) FROM bookingsB
    GROUP BY race;
```

```
A|3608

B|459772

I|525

U|386

W|758500
```

---

Or,

```sql
SELECT UPPER(race), UPPER(e), COUNT(bookingNumber) FROM bookingsB
GROUP BY UPPER(race), UPPER(e);
```

```
A|H|32

A|N|3519

A|U|57

B|H|9627

B|N|449775

B|U|370

I|H|7

I|N|509

I|U|9

U|H|226

U|N|150

U|U|10

W|H|180735

W|N|577119

W|U|646
```

b. How do we know if someone is homeless?

The only way we can tell (a priori) is to see if the address field doesn't make sense. For instance, if the address is NULL or "HOMELESS", etc. One could then write a script (which would probably take a while to get working correctly) that checks the county property appraiser's website to see if the property is in a residential neighborhood. If the propery is not, it's a good chance that the person is homeless.

Lastly, the way that we're going to take advantage of today, is if you somehow happen upon a list (or table) of known homeless shelters and other addresses commonly used by homeless people, but not used by non-homeless people, you can check the arrest's address against this list.

c. For people who are homeless, how many arrests are there for each charge?

```
CREATE TABLE chargeNBooking AS
    SELECT charge, bookingNumber, (UPPER(address) IN
                                   (SELECT UPPER(address) FROM homeless_addresses))AS homeless
        FROM bookingsB;

INSERT INTO chargeNBooking
      SELECT charge, bookingNumber, (SELECT homeless FROM chargeNBooking AS c
                                    WHERE bookingNumber = c.bookingNumber) AS homeless
        FROM booking_addl_charge;
```

```
TRESPASS ON PROP OTHER THAN STRUCTURE OR C|5856
POSSESSION OF COCAINE|5843
POSSESSION OF OPEN CONTAINER|4482
TRES. ON PROP. OTHER THAN STRUCT. OR CONVE|2456
DRIVING UNDER THE INFLUENCE|2289
NO VALID DRIVER|2249
MANUFACTURE.DIST.DISPENSE.POSSES CON SUB-|2193
PETIT THEFT ($100 OR LESS)|2167
POSSESSION OF CANNABIS LESS THAN 20 GRAMS|1772
BATTERY (DOMESTIC VIOLENCE)|1770
```

d. What do homeless people get charged with in greater proportion than others?

```
SELECT charge, count FROM homelessCharges AS h
  WHERE count > (SELECT count FROM nonHomelessCharges AS n
                  WHERE n.charge = h.charge)
  ORDER BY count DESC LIMIT 10;
```

```
SOLICITING ON EDGE OF ROAD|603
SOLICITATION AND DISTRIBUTION ON PUBLIC RO|266
UNLAWFUL CAMPING|166
UNLAWFUL USE OF STATE ROAD RIGHT OF WAY|66
UNLAWFUL SOLICITING IN PROHIBITED ZONE OR|14
TRAIN RIDING|10
FURNISHING INTOXICANTS TO HABITUAL DRUNKAR|3
DRIVING UNDER INFL|2
HIJACKING AIRCRAFT - FEDERAL|2
PUBLIC MUTILATION OF FLAG|2
```

---

e. What is the average stay in jail?

```sql
SELECT AVG(julianday(releaseDate) - julianday(arrestDate)) FROM bookingsA;
```

25.0 (approximately)

f. Where do arrests take place?

The arrest location was not included in the dataset, making this problem particularly difficult. (I have a dry sense of humor).

g. How many arrests are there per year for “possession of open container”, homeless vs. not.

```sql
SELECT count(bookingNumber) AS arrests, substr(arrestDate,1,4) AS year FROM bookingsA
  WHERE bookingNumber IN (SELECT bookingNumber FROM homelessCharges
                                WHERE charge LIKE "POSSESSION OF OPEN CONTAINER")
  GROUP by substr(arrestDate,1,4)
  ORDER by year
  LIMIT 10;
```

---

```
2|1947
4|1950
1|1951
1|1952
2|1955
2|1957
2|1958
2|1959
3|1960
1|1961
```

---

h. How old was each homeless person when first arrested?

```sql
SELECT name, MIN(julianday(arrestDate) - julianday(DOB))/365 AS age FROM bookingsA
    WHERE bookingNumber IN (SELECT bookingNumber FROM homelessCharges)
    GROUP BY name
    SORT BY age DESC
    LIMIT 10;
```

---

```
AARON,ANDREW LEE |42.9945205479452
AARON,LINDA LEE |42.1287671232877
ABADIA,MARIA DELPILAR |34.0027397260274
ABALA,GORGE NMN |26.7013698630137
ABALOS,DAVID O |44.827397260274
ABARCA,MARCELINO |22.4383561643836
ABARCA,SAMUEL |22.6328767123288
ABARCADESALMERON,JOBANI |30.1890410958904
ABBATIELLO,JOHN JOSEPH |46.5616438356164
ABBATT,ROBERT HAVEN |41.8027397260274
```