Exploring a Database (and exposing data)

Problem Sets 1 are due before midnight tonight in both sections.

Those in section 002 might want to access Jupyter today to follow along.

“Mommy wants you to know where your food comes from.”
Welcome Comp521-001!

What I know...

1) Professor Bishop fell ill and called off class last Thursday
2) He has since been hospitalized and expects to be incapacitated for AT LEAST this entire week.
3) Get problem set #1 turned in tonight

What I plan to do…

1) Use the next two lectures to sync up the two sections.
2) For today, assume that neither course has changed. I will be examining both courses syllabi and infrastructure

What if…

1) If Professor Bishop does not return, we will merge the two sections. I will also address discrepancies (grading, # of midterms, etc. …). Changes for both sections.
2) I hope to know more by Thursday
NCCOVID19.db from PS #2

You might want to...
- Open your PS #2
- Create a new cell to open the NCCOVID.db database
Simple Questions

❖ How many NC counties?

```sql
SELECT COUNT(*)
FROM County
```

❖ How many hospitals?

```sql
SELECT COUNT(*)
FROM Hospital
```

❖ How many races are used in the demographic data?

```sql
SELECT DISTINCT race
FROM Demographics
```

❖ How many years worth of demographics data are included.

```sql
SELECT DISTINCT year
FROM Demographics
```
More useful queries

❖ What were the 10 NC counties with the most COVID-19 confirmed cases on 2020-08-20?

```
SELECT C.name, V.date, V.cases
FROM County C, Covid19 V
WHERE C.fips=V.fips AND date="2020-08-20"
ORDER BY cases DESC
LIMIT 10
```

❖ What are the metropolitan areas are in each of these counties?
❖ On what days did Orange county have its most reported cases?
Another query

- Find the breakdown of North Carolina's 2020 population by race

  ```sql
  SELECT D.race, SUM(D.count) 
  FROM Demographics D 
  WHERE D.year=2020 
  GROUP BY D.race
  ```

- Find the total 2020 populations by county

- List the all NC counties that in 2020 are not majority "white"
Queries Continued

❖ How many hospitals are in each county?

```
SELECT C.name, COUNT(*)
FROM County C, Hospital H, HospitalCounty HC
WHERE C.fips=HC.fips AND H.hid=HC.hid AND HC.incounty=1
GROUP BY C.fips
ORDER BY COUNT(*) DESC
```

❖ Modify the query above to include every county? (Hint: think LEFT JOIN)
Where are college age folks?

- List the population for the age group from 18 to 24 in every NC county in 2020
  
  ```sql
  SELECT C.name, SUM(D.count)
  FROM County C, Demographics D
  WHERE C.fips=D.fips AND D.agelo >= 18 AND D.agelo <= 24
  AND year=2020
  GROUP BY D.fips
  ORDER BY SUM(D.count) DESC
  ```

- Do the same query, but also give it as a ratio of the total population
How many hospital and hospital beds are in each county?

SELECT C.name, Count(H.hid), SUM(H.beds)
FROM County C, Hospital H, HospitalCounty HC
WHERE C.fips=HC.fips AND H.hid=HC.hid AND HC.incounty=1
GROUP BY C.fips
ORDER BY SUM(H.beds) DESC
Can you trust the data?

❖ How many hospitals have no beds?

Actually, this isn't true. I grabbed the hospital data from a website where the hospitals voluntarily reported.
Cleaning the data

❖ Here is how we fix it

```
UPDATE Hospital SET beds=185 WHERE hid=1060
```

❖ Now, recompute the number of hospital beds in each county?

```
SELECT C.name, Count(H.hid), SUM(H.beds)
FROM County C, Hospital H, HospitalCounty HC
WHERE C.fips=HC.fips AND H.hid=HC.hid AND HC.incounty=1
GROUP BY C.fips
ORDER BY SUM(H.beds) DESC
```
Time to COMMIT

❖ No change is permanently made to a data base until it is committed!

❖ Now commit

❖ Now if we restart iSQL the fix will remain.

❖ Lack of COMMITs is a major source of bugs
  ▪ "locked" database
  ▪ loss of data

“Why are you so afraid to commit? Do you really think you’re going to meet someone better than me?”
Errors, errors, everywhere

❖ Dirty data is everywhere
  - Some hospitals have closed
  - New ones added
  - The list of counties "served" by hospitals is incomplete
  - It takes a it of sleuthing to figure them out

❖ However, we need to be able to write correct queries in the presence of imperfect data

❖ This is the challenge and learning objective of your next problem set
  - I have fixed many errors that appear in the version of NCCOVID19.db that you have
  - I may have also probably introduced dozens more

❖ Your HW queries need to work correctly on my instance as well as yours.
Next Time

❖ Embedding SQL queries within a traditional programming language