CS564 Fall '14: SQL Assignment

Individual Assignment

Due: Tuesday, October 14th by 11:30 PM

---

1 Introduction

In this assignment you will use **SQLite3** to execute a few SQL queries. You can find documentation on SQLite3 at here. SQLite3 is not as functional as PostgreSQL or MySQL or the commercial relational DBMSes, but it is much easier to use and query. That is why there are rumoured to be many more installations of SQLite than PostgreSQL or any other traditional DBMS. Also, you can easily install SQLite3 on your own machine, but make sure your final SQL queries run on SQLite3 installed on the **CS mumble machines**.

For this assignment, we will use the US Census Population Estimate dataset, which can be found here. We have downloaded and cleaned this dataset and inserted it into a sqlite database called *census.db* that can be found in the directory called *Pop_Estimate_Dataset* in this assignment directory.

There is a layout file for each table in the database. The table layout is helpful in understanding what each attribute means. Section 3 of this document gives you details of the schema of the database.

2 Query the Database

Your task is to write the following queries in separate files, as described below:

File	Query	Output Columns
Query1.sql	List all states and their values for 2011 housing estimate (**HUEST_2011**)	(State Code, Housing Estimate)
Query2.sql	Find the states which have the number of births in 2012 (**BIRTS2012**) to be greater than 80,000	(State Code, Births)
Query3.sql	For each state, print its name and the net migration into the state for the year 2012 (**NETMIG2012**)	(State Name, Net Migration)
Query4.sql	Find the number of counties in each state and print it with the name of the state	(State Name, Count of Counties)
Query5.sql	Display names of all "metropolitan statistical areas"(**LSAD**) and their corresponding migration for the period between June 2010 and June 2011 (**NETMIG2011**)	(Area Name, Net Migration)
Query6.sql	Find each state, print the population estimate in the year 2011 (**POPESTIMATE2011**) of all women aged between 21 and 45 (both inclusive)	(Age, Population)
Query7.sql	For each division, print the name of the division 1∼9 (**DIVISION_DESC**) and a list of all states that belong to that region concatenated into one string that is separated by commas (take a look at the sqlite documentation for the appropriate aggregate function)	(Division Name, State List)

Query8.sql	For each division, find the name of the state that has the greatest increase in housing estimates for the year 2011 (**HUEST_2011**) as compared to housing estimates for the year 2010 (**HUEST_2010**)	(Division Name, State Name, Increase)
Query9.sql	For each states, display the percentage of the states estimated 2011 population (**POPESTIMATE2011**) that is above the age of 21 ($\geq$ 21)	(State Name, Percentage)
Query10.sql	Find the states for which the ratio of the "2011 resident total population estimates"(**POPESTIMATE2011**) and "2011 housing estimates base"(**HUEST_2011**) is below the nationwide ratio. The ratios should be compared in floating point number, not in integer	(State Name)
Query11.sql	For each state, find the name of the county that has the smallest increase in the population between 2011 and 2012 (**NPOPCHG_2012**). Look at counties individually and not the state as a whole. Display name of state and county and population change in lexicographic order of "Division_Desc". For states with same division, sort in decreasing order of the 2012 population estimate (**POPESTIMATE2012**) of the county	(State Name, County, Population Change)
Query12.sql	For all age groups, calculate the absolute difference of 2011 population estimate (**POPESTIMATE2011**) and 2010 population estimate (**POPESTIMATE2010**). Also, display the indicator that indicates if the population has increased or decreased, or remained the same. The value of indicator should be one of {increased, same, decreased}	(Age, Difference, Indicator)

—

The sqlite file itself is in binary and does not convey any information about the tables present in it. You may want to use sqlitestudio. This allows you to visualize the tables and run queries on the data which might be useful in the above task.

If you cant write any one of the queries above, provide an explanation of the features that are lacking in SQLite3 that prevents you from being able to write that query. Write these explanations in a file called Queries-readme.txt. To test whether or not your queries work as expected, you can use sqlitestudio or the command below.

Note: To check the query output, we will run the following command on our copy of *census.db*:

```
$ sqlite3 census.db < Query3.sql
```

Thus, your text file can have multiple queries, but it must only output your final desired output, and clean up any temporary tables that it creates. You should try to minimize the number of distinct SQL query blocks in each query file – **try to write a single query for each of the queries above, whenever possible**.

Copy each of your query's result into a separate file called result[1-12].txt, i.e., the word result followed by the query number and .txt. For example, if we want to get the results of Query 5, we would execute the following:

```
$ sqlite3 census.db < Query5.sql > result5.txt
```

The above command can be repeated for each query with the proper query number substituted in place to get all 12 result files.

3 Schema for the Population Database

Here is a list of tables that are present in the database and the corresponding files in which you can find functional details such as the semantics of each column of the table.

Table Name	Where to Find the Functional Details About the Table
POP_ESTIMATE_NATION_STATE_PR	POP_ESTIMATE_NATION_STATE_PR.pdf
POP_ESTIMATE_STATE_COUNTY	POP_ESTIMATE_STATE_COUNTY.pdf
POP_ESTIMATE_CITIES_TOWNS	POP_ESTIMATE_CITIES_TOWNS.pdf
POP_ESTIMATE_STATE_AGE_SEX_RACE_ORIGIN	POP_ESTIMATE_STATE_AGE_SEX_RACE_ORIGIN.pdf
POP_ESTIMATE_PR_SEX_AGE	POP_ESTIMATE_PR_SEX_AGE.pdf
POP_ESTIMATE_PR_MUNICIPIOS	POP_ESTIMATE_PR_MUNICIPIOS.pdf
POP_ESTIMATE_PR_MUN_SEX_AGE	POP_ESTIMATE_PR_MUN_SEX_AGE.pdf
POP_ESTIMATE_METRO_MICRO	POP_ESTIMATE_METRO_MICRO.htm
HU_UNIT_STATE_LEVEL	HU_UNIT_STATE_LEVEL.pdf

The above mentioned pdf and htm files are present along with the data files in the *Pop_Estimate_Dataset* directory.

We have created six more tables listed below. The above population tables have few columns such as Division, Region, Origin, Race, Sex, Sumlev that have numeric values in it. The meaning for those numeric values can be obtained from these separate tables below. The schemas for these tables are provided below along with the schemas for other tables.

Table Name
REGION
DIVISION
RACE
SEX
ORIGIN
SUMLEV

The file layout for each Population table does not give any information about the Data Type and any Constraints present on the table. It can be referred to know the functional details as to what information is present in each table and what each column in the table means.

We have listed the data type and constraints (Primary and Foreign Key) for each table below. The attribute with value *FK* in the *Cons.* field is a foreign key to the primary key of the table with the same name. As you can see below, there are cases where the same attribute is both a primary and foreign key. This might be useful when it comes to writing the actual queries.

REGION

Column	Type	Cons.
REGION_CD	Number	PK
REGION_DESC	String	

SEX

Column	Type	Cons.
SEX_CD	Number	PK
SEX_DESC	String	

DIVISION

Column	Type	Cons.
DIVISION_CD	Number	PK
DIVISION_DESC	String	

RACE

Column	Type	Cons.
RACE_CD	Number	PK
RACE_DESC	String	

ORIGIN

Column	Type	Cons.
ORIGIN_CD	Number	PK
ORIGIN_DESC	String	

SUMLEV

Column	Type	Cons.
SUMLEV_CD	Number	PK
SUMLEV_DESC	String	

POP_ESTIMATE_NATION_STATE_PR

Column	Type	Cons.
SUMLEV	Number	FK
REGION	Number	FK
DIVISION	Number	FK
STATE	Number	
NAME	String	PK
CENSUS2010POP	Number	
ESTIMATESBASE2010	Number	
POPESTIMATE2010	Number	
POPESTIMATE2011	Number	
POPESTIMATE2012	Number	
NPOPCHG_2010	Number	
NPOPCHG_2011	Number	
NPOPCHG_2012	Number	
BIRTHS2010	Number	
BIRTHS2011	Number	
BIRTHS2012	Number	
DEATHS2010	Number	
DEATHS2011	Number	
DEATHS2012	Number	
NATURALINC2010	Number	
NATURALINC2011	Number	
NATURALINC2012	Number	
INTERNATIONALMIG2010	Number	
INTERNATIONALMIG2011	Number	
INTERNATIONALMIG2012	Number	
DOMESTICMIG2010	Number	
DOMESTICMIG2011	Number	
DOMESTICMIG2012	Number	
NETMIG2010	Number	
NETMIG2011	Number	
NETMIG2012	Number	
RESIDUAL2010	Number	
RESIDUAL2011	Number	
RESIDUAL2012	Number	
RBIRTH2011	Real	
RBIRTH2012	Real	
RDEATH2011	Real	
RDEATH2012	Real	
RNATURALINC2011	Real	
RNATURALINC2012	Real	
RINTERNATIONALMIG2011	Real	
RINTERNATIONALMIG2012	Real	
RDOMESTICMIG2011	Real	
RDOMESTICMIG2012	Real	
RNETMIG2011	Real	
RNETMIG2012	Real	

POP_ESTIMATE_STATE_COUNTY

Column	Type	Cons.
SUMLEV	Number	FK
REGION	Number	FK
DIVISION	Number	FK
STATE	Number	PK
COUNTY	Number	PK
STNAME	String	
CTYNAME	String	
CENSUS2010POP	Number	
ESTIMATESBASE2010	Number	
POPESTIMATE2010	Number	
POPESTIMATE2011	Number	
POPESTIMATE2012	Number	
NPOPCHG_2010	Number	
NPOPCHG_2011	Number	
NPOPCHG_2012	Number	
BIRTHS2010	Number	
BIRTHS2011	Number	
BIRTHS2012	Number	
DEATHS2010	Number	
DEATHS2011	Number	
DEATHS2012	Number	
NATURALINC2010	Number	
NATURALINC2011	Number	
NATURALINC2012	Number	
INTERNATIONALMIG2010	Number	
INTERNATIONALMIG2011	Number	
INTERNATIONALMIG2012	Number	
DOMESTICMIG2010	Number	
DOMESTICMIG2011	Number	
DOMESTICMIG2012	Number	
NETMIG2010	Number	
NETMIG2011	Number	
NETMIG2012	Number	
RESIDUAL2010	Number	
RESIDUAL2011	Number	
RESIDUAL2012	Number	
GQESTIMATESBASE2010	Number	
GQESTIMATES2010	Number	
GQESTIMATES2011	Number	
GQESTIMATES2012	Number	
RBIRTH2011	Real	
RBIRTH2012	Real	
RDEATH2011	Real	
RDEATH2012	Real	
RNATURALINC2011	Real	
RNATURALINC2012	Real	
RINTERNATIONALMIG2011	Real	
RINTERNATIONALMIG2012	Real	
RDOMESTICMIG2011	Real	
RDOMESTICMIG2012	Real	
RNETMIG2011	Real	
RNETMIG2012	Real	

POP_ESTIMATE_METRO_MICRO

Column	Type	Cons.
CBSA	Number	
MDIV	Number	
STCOU	Number	
NAME	String	PK
LSAD	String	PK
CENSUS2010POP	Number	
ESTIMATESBASE2010	Number	
POPESTIMATE2010	Number	
POPESTIMATE2011	Number	
NPOPCHG_2010	Number	
NPOPCHG_2011	Number	
NATURALINC2010	Number	
NATURALINC2011	Number	
BIRTHS2010	Number	
BIRTHS2011	Number	
DEATHS2010	Number	
DEATHS2011	Number	
NETMIG2010	Number	
NETMIG2011	Number	
INTERNATIONALMIG2010	Number	
INTERNATIONALMIG2011	Number	
DOMESTICMIG2010	Number	
DOMESTICMIG2011	Number	
RESIDUAL2010	Number	
RESIDUAL2011	Number	

HU_UNIT_STATE_LEVEL

Column	Type	Cons.
SUMLEV	Number	FK
STATE	Number	FK
REGION	Number	PK
DIVISION	Number	FK
STNAME	String	
HUCENSUS2010	Number	
HUESTBASE2010	Number	
HUEST_2010	Number	
HUEST_2011	Number	

POP_ESTIMATE_PR_MUN_SEX_AGE

Column	Type	Cons.
SUMLEV	Number	FK
MUNICIPIO	Number	PK
NAME	String	
YEAR	Number	PK
POPESTIMATE	Number	
POPEST_MALE	Number	
POPEST_FEM	Number	
UNDER5_TOT	Number	
UNDER5_MALE	Number	
UNDER5_FEM	Number	
AGE513_TOT	Number	
AGE513_MALE	Number	
AGE513_FEM	Number	
AGE1417_TOT	Number	
AGE1417_MALE	Number	
AGE1417_FEM	Number	
AGE1824_TOT	Number	
AGE1824_MALE	Number	
AGE1824_FEM	Number	
AGE16PLUS_TOT	Number	
AGE16PLUS_MALE	Number	
AGE16PLUS_FEM	Number	
AGE18PLUS_TOT	Number	
AGE18PLUS_MALE	Number	
AGE18PLUS_FEM	Number	
AGE1544_TOT	Number	
AGE1544_MALE	Number	
AGE1544_FEM	Number	
AGE2544_TOT	Number	
AGE2544_MALE	Number	
AGE2544_FEM	Number	
AGE4564_TOT	Number	
AGE4564_MALE	Number	
AGE4564_FEM	Number	
AGE65PLUS_TOT	Number	
AGE65PLUS_MALE	Number	
AGE65PLUS_FEM	Number	
AGE85PLUS_TOT	Number	
AGE85PLUS_MALE	Number	
AGE85PLUS_FEM	Number	
MEDIAN_AGE_TOT	Number	
MEDIAN_AGE_MALE	Number	
MEDIAN_AGE_FEM	Number	

POP_ESTIMATE_CITIES_TOWNS

Column	Type	Cons.
SUMLEV	Number	FK
STATE	Number	PK
COUNTY	Number	PK
PLACE	Number	PK
COUSUB	Number	PK
CONCIT	Number	PK
NAME	String	
STNAME	String	
CENSUS2010POP	Number	
ESTIMATESBASE2010	Number	
POPESTIMATE2010	Number	
POPESTIMATE2011	Number	

POP_ESTIMATE_PR_MUNICIPIOS

Column	Type	Cons.
SUMLEV	Number	FK
MUNICIPIO	Number	PK
NAME	String	
ESTIMATESBASE2010	Number	
POPESTIMATE2010	Number	
POPESTIMATE2011	Number	
NPOPCHG_2010	Real	
NPOPCHG_2011	Real	
PPOPCHG_2010	Real	
PPOPCHG_2011	Real	
SRANK_ESTBASE2010	Number	
SRANK_POPEST2010	Number	
SRANK_POPEST2011	Number	
SRANK_NPCHG2010	Number	
SRANK_NPCHG2011	Number	
SRANK_PPCHG2010	Number	
SRANK_PPCHG2011	Number	

POP_ESTIMATE_STATE_AGE_SEX_RACE_ORIGIN

Column	Type	Cons.
SUMLEV	Number	FK
REGION	Number	FK
DIVISION	Number	FK
STATE	Number	PK
SEX	Number	PK, FK
ORIGIN	Number	PK, FK
RACE	Number	PK, FK
AGE	Number	PK
CENSUS2010POP	Number	
ESTIMATESBASE2010	Number	
POPESTIMATE2010	Number	
POPESTIMATE2011	Number	

POP_ESTIMATE_PR_SEX_AGE

Column	Type	Cons.
SUMLEV	Number	FK
STATE	Number	
NAME	String	
SEX	Number	PK, FK
AGE	Number	PK
CENSUS2010POP	Number	
ESTIMATESBASE2010	Number	
POPESTIMATE2010	Number	
POPESTIMATE2011	Number	

4 Submission

Create only the following files for this assignment: Query[1-12].sql, result[1-12].txt, Queries-readme.txt. Compress these files into a `tar.gz` package. The man pages for `tar` should be able to helf you with this. The name of your compressed package should be `<your_netid>_A2.tar.gz`. For example, if your UW-netid is `johndoe@wisc.edu`, your compressed package should be called `johndoe_A2.tar.gz`. Please use comments in all your SQL files to better explain what you are trying to accomplish in each query. Please upload this tar file to the `Assignment 2 Submission` Dropbox folder in learn@uw.

Acknowledgements

This assignment is based in part on the assignment by Prof. Jignesh Patel especially parts concerning cleaning up of the data and separating them into individual files.