

Census Database

Adam Stammer, Isaac Plevak





Roles

Data Acquisition - Adam

- Obtaining and formatting data for import

Database Design - Isaac

- Planning and creation of the database



Summary

Why Census Bureau Data?

- Real life data that we wouldn't have to manually create
- Census data is easy to acquire
- Data was in XLS format which allowed easier implementation with SQL
- Legitimate questions that can be answered based on this data



Data Acquisition

XLS to CSV via Python script

- Used Pandas and Numpy Libraries

Batch script to concatenate separate csv files into one file

- Much easier to import one file than 50

Source Data

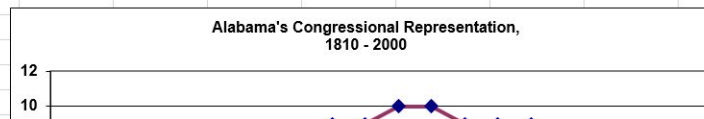
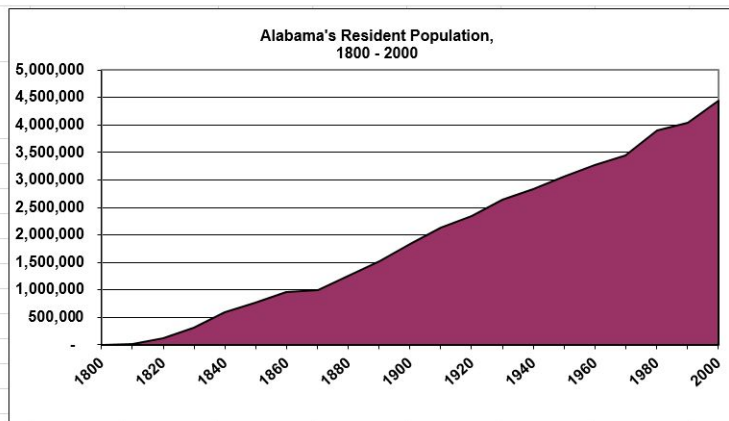
U.S. Census Bureau

Resident Population and Apportionment of the U.S. House of Representatives



Alabama

Year	Resident Population	Number of Representatives
2000	4,447,100	7
1990	4,040,587	7
1980	3,894,025	7
1970	3,444,354	7
1960	3,266,740	8
1950	3,061,743	9
1940	2,832,961	9
1930	2,646,248	9
1920	2,348,174	10
1910	2,138,093	10
1900	1,828,697	9
1890	1,513,401	9
1880	1,262,505	8
1870	996,992	8
1860	964,201	6
1850	771,623	7
1840	590,756	7





Formatted Data

year	population	number_of_reps	state_Name
1980	3894025	7	Alabama
1970	3444354	7	Alabama
1960	3266740	8	Alabama
1950	3061743	9	Alabama
1940	2832961	9	Alabama
1930	2646248	9	Alabama
1920	2348174	10	Alabama
1910	2138093	10	Alabama
1900	1828697	9	Alabama
1890	1513401	9	Alabama
1880	1262505	8	Alabama
1870	996992	8	Alabama
1860	964201	6	Alabama
1850	771623	7	Alabama
1840	590756	7	Alabama
1830	309527	5	Alabama
1820	127901	3	Alabama
1810	9046	1	Alabama



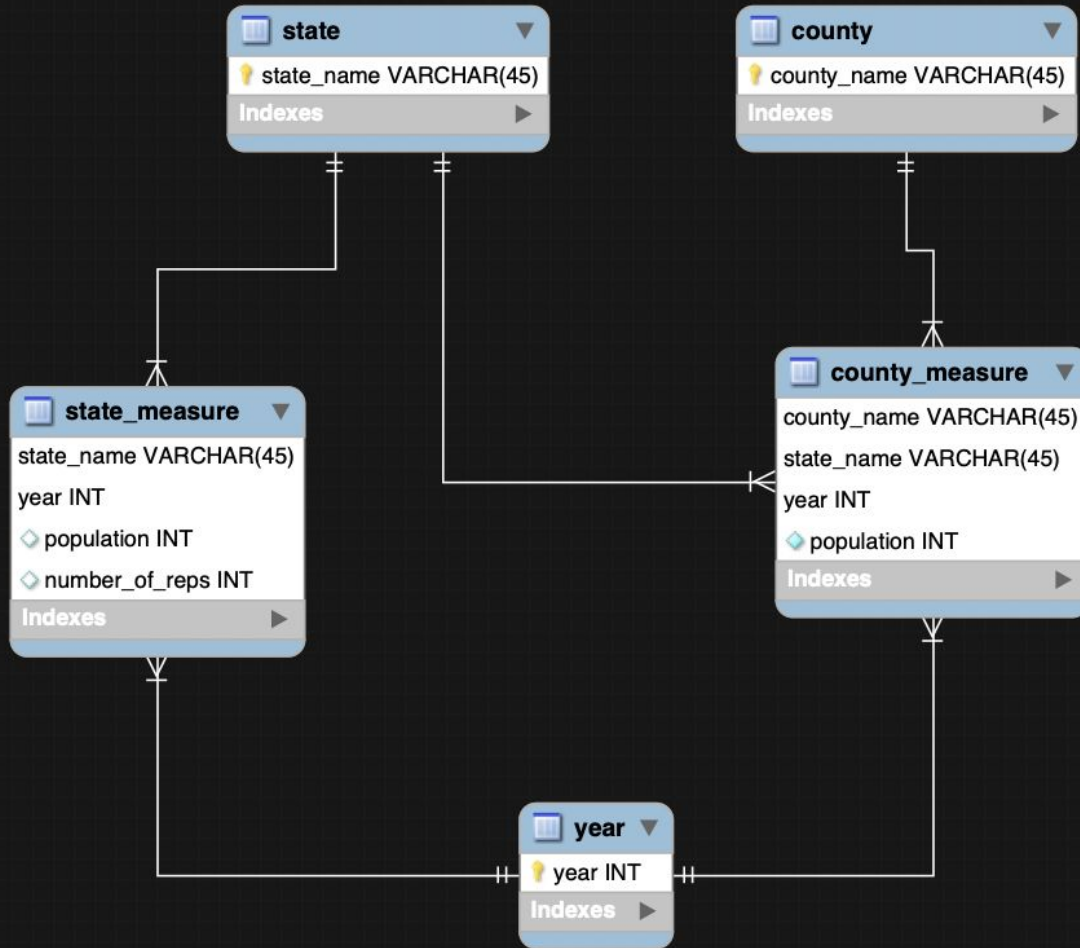
Design

Original design included region data and city data.

Current design no longer has city data but has county data.

Region population is accessed by adding the population of each state together of a given region. This is a great opportunity to use views.

Entity Relationship Diagram



state is a separate table from state_measure.

year has its own table.

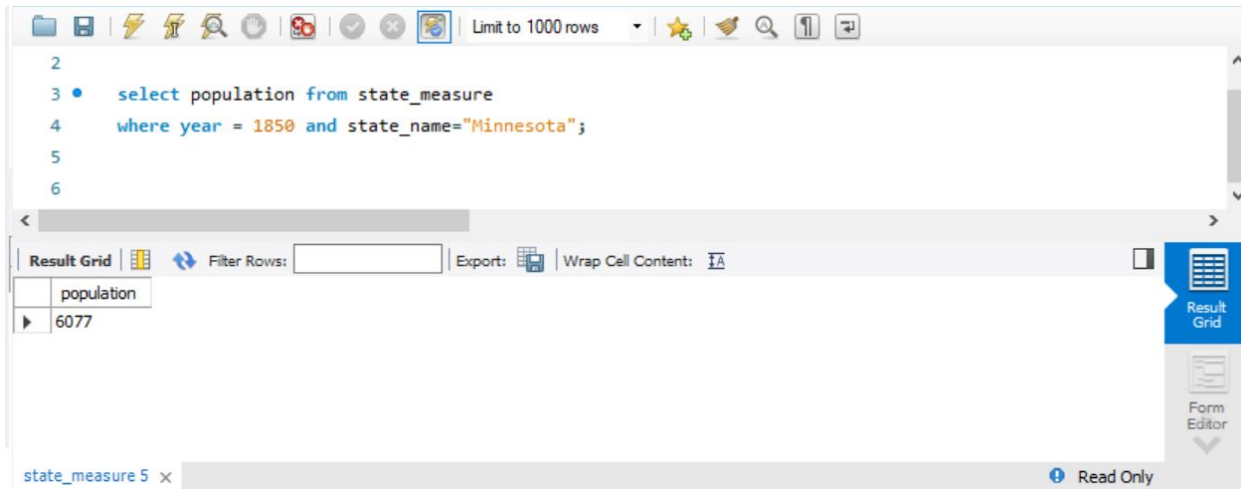
state_measure used foreign keys to reference this information.

This helps eliminate redundant information from being stored.

Use Cases



What was the Minnesota state population in 1850?



The screenshot shows a data query interface. At the top, there is a toolbar with various icons and a dropdown menu set to "Limit to 1000 rows". Below the toolbar is a text area containing a SQL query:

```
2  
3 • select population from state_measure  
4   where year = 1850 and state_name="Minnesota";  
5  
6
```

Below the query area is a section labeled "Result Grid". It contains a table with one row of data:

population
6077

Below the table, there is a "Filter Rows:" input field, an "Export:" button, and a "Wrap Cell Content:" checkbox. On the right side of the interface, there are buttons for "Result Grid" and "Form Editor". At the bottom, there is a status bar that says "state_measure 5 x" and "Read Only".

Which counties in Minnesota had a greater population than 250,000 in 2000?

The screenshot shows a data analysis tool interface. At the top, there is a toolbar with various icons for file operations, search, and execution. Below the toolbar, a SQL query is entered in a text area:

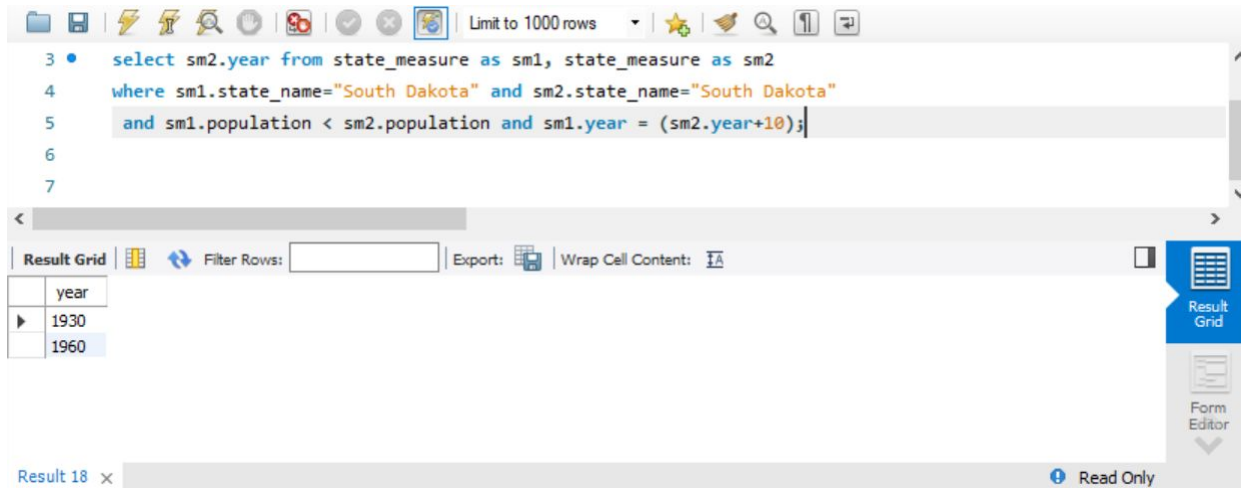
```
2  
3 • select county_name from county_measure  
4   where year = 2000 and population > 250000;  
5  
6
```

Below the query editor, there is a "Result Grid" section. It includes a "Filter Rows" input field, an "Export" button, and a "Wrap Cell Content" checkbox. The results are displayed in a table with the following data:

county_name
Anoka
Dakota
Hennepin
Minnesota
Ramsey

On the right side of the interface, there are buttons for "Result Grid" and "Form Editor". At the bottom, a status bar indicates "county_measure 15" and "Read Only".

During which decades did the South Dakota state population decline?



The screenshot shows a data analysis tool interface. At the top, there is a toolbar with various icons. Below the toolbar, a SQL query is displayed in a text area:

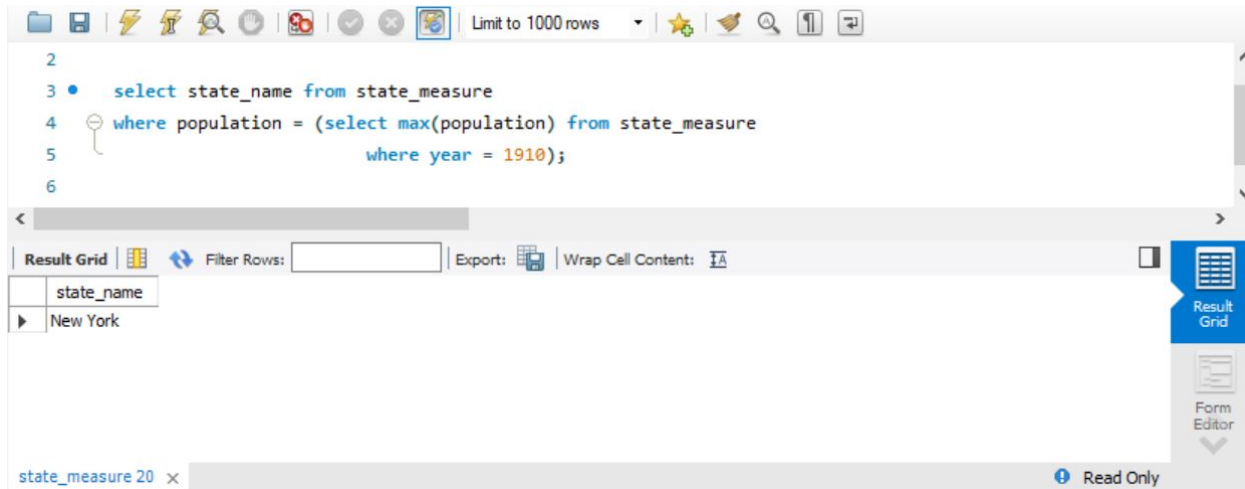
```
3 • select sm2.year from state_measure as sm1, state_measure as sm2
4 where sm1.state_name="South Dakota" and sm2.state_name="South Dakota"
5 and sm1.population < sm2.population and sm1.year = (sm2.year+10);
6
7
```

Below the query, there is a section for the results. It includes a "Result Grid" button, a "Filter Rows" input field, and an "Export" button. The "Result Grid" is currently displaying the following data:

year
1930
1960

At the bottom of the interface, there is a status bar that reads "Result 18 x" and "Read Only".

In 1910, which state had the greatest population?



The screenshot shows a data analysis tool interface. At the top, there is a toolbar with various icons and a dropdown menu set to "Limit to 1000 rows". Below the toolbar is a text area containing a SQL query:

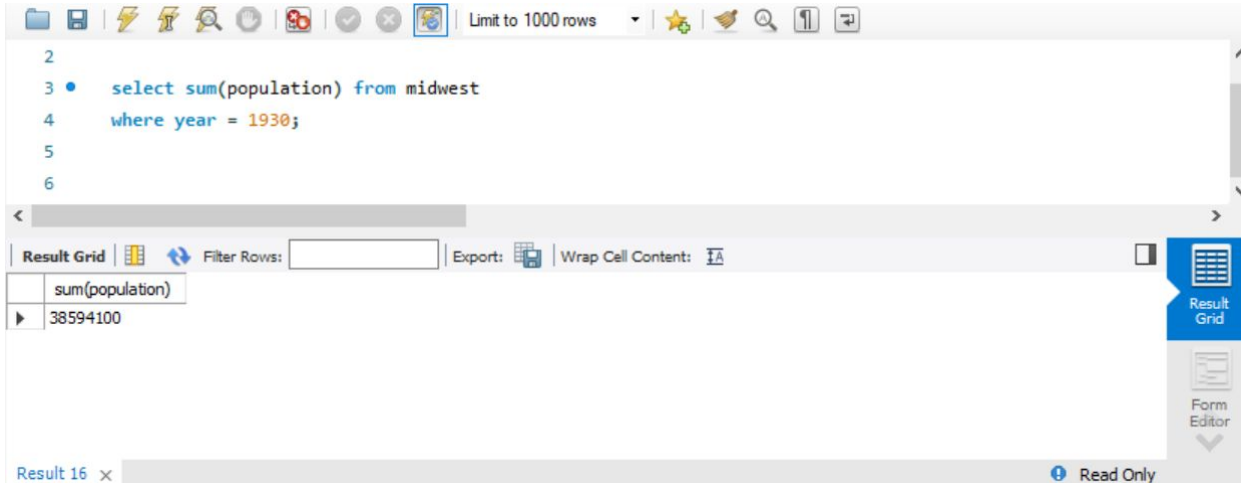
```
2  
3 • select state_name from state_measure  
4 where population = (select max(population) from state_measure  
5                       where year = 1910);  
6
```

Below the query editor is a section labeled "Result Grid". It includes a "Filter Rows:" input field, an "Export:" button, and a "Wrap Cell Content:" checkbox. The results are displayed in a table with one row:

state_name
New York

On the right side of the interface, there are buttons for "Result Grid" and "Form Editor". At the bottom, there is a tab labeled "state_measure 20" and a "Read Only" indicator.

What was the total population of the midwest region in 1930?



The screenshot shows a SQL query editor interface. The query is:

```
select sum(population) from midwest
where year = 1930;
```

Below the query, the results are displayed in a table:

sum(population)
38594100

The interface includes a toolbar at the top with various icons and a "Limit to 1000 rows" dropdown. On the right side, there are buttons for "Result Grid" and "Form Editor". At the bottom, it says "Result 16 x" and "Read Only".



Questions?



Sources

<https://www.census.gov>