Ian Sargent
CS 2500
Project 1: Create A Database

## Final Deliverables

Original Data Source

Original Datasets

R Script to Clean the Datasets

Cleaned Datasets

The Created Database

**Final Tables(3):**

**Employment**(State, County, Civ_Labor_Force_2023, Med_Household_Income, Unemployment_Rate_2023)

**Poverty**(State, County, Poverty_All_Ages_2023, Poverty_Pct_All_Ages_2023, Poverty_0to17_2023, Rural_Urban_Code_2023)

**Population**(State, County, Census_Pop_2020, Population_Estimate_2023, Births_2023, Deaths_2023)

**Note**: Connecticut counties have been excluded from the Employment and Population Table. Some combined datasets that make up these tables divided CT into planning regions, whereas others divided CT into its formal counties. Since these areas are geographically different, joining them would create inaccurate data.
U.S.D.A. … DO BETTER!

## Statement of Interest

I believe that national data such as the data contained in this project should be publicly accessible and thoroughly documented since they are made possible with taxpayer money. Leveraging this available information about the U.S. on the county level allows for a better grasp of regional demographic and economic trends. Accessible government data empowers researchers, policymakers, and the public to make informed decisions based on empirical evidence. I've decided to practice my right to this information with this project, downloading tables directly from the USDA website to analyze them. There are many surprising results within these tables about counties across the United States. This project highlights States with the lowest median household income, counties with the lowest unemployment rates, and areas with the fastest growing populations in the country. People must remain knowledgeable about these trends to inform policies and make better decisions for our country. By querying this database, I aim to shed light on disparities and growth patterns that might go unnoticed. Transparency and accessibility are key to a well-informed society. I also aim to develop my data-cleaning skills in R and database querying skills in SQL.

## Queries (SQL and Relational Algebra)

---

**Question 1**: *What is the estimated population of each Texas county in 2023?*

**Relational Algebra**:

```
π County,
Population_Estimate_2023(σ State =
'TX'(Population))
```

**SQLite Query**:

```
SELECT County,
Population_Estimate_2023 FROM
Population
WHERE State = 'TX';
```

| | County | Population_Estimate_2023 |
|---|---|---|
| 1 | Anderson County, TX | 57736 |
| 2 | Andrews County, TX | 18664 |
| 3 | Angelina County, TX | 87319 |
| 4 | Aransas County, TX | 25374 |
| 5 | Archer County, TX | 9029 |
| 6 | Armstrong County, TX | 1832 |
| 7 | Atascosa County, TX | 51784 |
| 8 | Austin County, TX | 31677 |
| 9 | Bailey County, TX | 6672 |
| 10 | Bandera County, TX | 22637 |
| 11 | Bastrop County, TX | 110778 |
| 12 | Baylor County, TX | 3463 |
| 13 | Bee County, TX | 30850 |
| 14 | Bell County, TX | 393193 |
| 15 | Bexar County, TX | 2087679 |

**Table Output**:

*254 Rows*

**Interpretation**:

This table lists all 254 Texas counties and their estimated 2023 populations. Although little statistical analysis results from this table, we do see large population discrepancies among certain counties. This shows the diversity of county population sizes in Texas.

---

**Question 2**: *What are the top ten counties with the highest unemployment rates in 2023?*

**Relational Algebra**:

π County, Unemployment_Rate_2023 TOP 10(τ Unemployment_Rate_2023 DESC (Employment))

**SQLite Query**:

```
SELECT County, Unemployment_Rate_2023
FROM Employment
ORDER BY Unemployment_Rate_2023 DESC
LIMIT 10;
```

| | County | Unemployment_Rate_2023 |
|---|---|---|
| 1 | Imperial County, CA | 17.3 |
| 2 | Kusilvak Census Area, AK | 15.1 |
| 3 | Yuma County, AZ | 13.2 |
| 4 | Colusa County, CA | 13 |
| 5 | Jefferson County, MS | 11.9 |
| 6 | Tulare County, CA | 10 |
| 7 | Luna County, NM | 10 |
| 8 | Magoffin County, KY | 9.8 |
| 9 | Bethel Census Area, AK | 9.7 |
| 10 | Northwest Arctic Borough, AK | 9.6 |

**Table Output**:

*10 Rows*

**Interpretation**:

This table lists the counties with the highest unemployment rates in the country. Three of the ten counties are in California, and three are in Arkansas. No apparent regional trend in this data emerges; however, Imperial County, CA, has the highest 2023 unemployment rate in the U.S. at 17.3%. It is located on the southeast tip of the state. This is attributed to the isolated geographic nature of the land and its reliance on seasonal agricultural work.

---

**Question 3**: *List the counties with birth-to-death ratios greater than 1. List the highest ratios at the top of the table.*

**Relational Algebra**:

```
π County, (Births_2023 /
Deaths_2023) as BD_Ratio (τ
BD_Ratio DESC (σ (Births_2023 /
Deaths_2023 > 1.0) (Population)))
```

**SQLite Query**:

```
SELECT County, ROUND(Births_2023 *
1.0/Deaths_2023, 2) AS BD_ratio
FROM Population
WHERE BD_ratio > 1.0
ORDER BY BD_ratio DESC;
```

| | County | BD_ratio |
|---|---|---|
| 1 | Madison County, ID | 5.29 |
| 2 | Chattahoochee County, GA | 4.58 |
| 3 | Grant County, NE | 4.0 |
| 4 | Utah County, UT | 3.9 |
| 5 | Clark County, ID | 3.8 |
| 6 | Geary County, KS | 3.49 |
| 7 | Loudoun County, VA | 3.04 |
| 8 | Kusilvak Census Area, AK | 3.0 |
| 9 | Summit County, CO | 2.94 |
| 10 | Cache County, UT | 2.91 |
| 11 | Gaines County, TX | 2.9 |
| 12 | Harding County, SD | 2.83 |
| 13 | Eagle County, CO | 2.78 |
| 14 | Rockland County, NY | 2.7 |
| 15 | Prince William County, VA | 2.68 |

**Table Output**:

*960 Rows*

**Interpretation**:

The 960 counties listed above each have a positive birth-to-death ratio. The birth-to-death ratio is a strong indicator of a growing population. Listed at the top are the counties growing the fastest (ignoring migration patterns). Madison County in Idaho has the highest birth-to-death ratio, with approximately five births for every death.

---

**Question 4**: *Which county with a rural-urban code of 1 has the highest percentage of people living in poverty in 2023?*

**Relational Algebra**:

π County, Poverty_Pct_All_Ages_2023 (τ Poverty_Pct_All_Ages_2023 DESC (σ Rural_Urban_Code = 1(Poverty)))

**SQLite Query**:

```
SELECT County, Poverty_Pct_All_Ages_2023 FROM Poverty
WHERE Rural_Urban_Code_2023 = 1
ORDER BY Poverty_Pct_All_Ages_2023 DESC
LIMIT 1;
```

**Table Output**:

*1 Row*

| | County | Poverty_Pct_All_Ages_2023 |
|---|---|---|
| 1 | Bronx County, NY | 27.7 |

**Interpretation**:

According to the USDA, a rural-urban code of 1 represents a "county in a metro area with a population of 1,000,000 or more." Out of those U.S. counties, the Bronx has the highest poverty rate for all ages, with 27.7% of residents living in poverty.

---

**Question 5**: *What is the civilian labor force participation rate for each state in 2023? (Hint: Participation_Rate = Labor_Force / Population_Estimate_2023)*

**Relational Algebra**:

```
τ Lab_Participation_Rate DESC
(γ e.State; SUM(e.Civ_Labor_Force_2023)/
SUM(p.Population_Estimate_2023) →
Lab_Participation_Rate (Employment ⋈
Employment.County = Population.County
Population))
```

**SQLite Query**:

```
SELECT e.State,
       Round(SUM(e.Civ_Labor_Force_2023) * 1.0
/ SUM(p.Population_Estimate_2023), 2) AS
Lab_Participation_Rate
FROM Employment e
JOIN Population p
ON e.County = p.County
GROUP BY e.State
ORDER BY Lab_Participation_Rate DESC;
```

| | State | Lab_Participation_Rate |
|---|---|---|
| 1 | DC | 0.59 |
| 2 | VT | 0.54 |
| 3 | NH | 0.54 |
| 4 | MN | 0.54 |
| 5 | MA | 0.54 |
| 6 | CO | 0.54 |
| 7 | WI | 0.53 |
| 8 | NE | 0.53 |
| 9 | ND | 0.53 |
| 10 | IA | 0.53 |
| 11 | WA | 0.52 |
| 12 | VA | 0.52 |
| 13 | UT | 0.52 |
| 14 | SD | 0.52 |
| 15 | RI | 0.52 |

**Table Output**:
*50 Rows*
*Connecticut (CT) has been excluded from this table due to the issues noted above.*

**Interpretation**:

The civilian labor force participation rates are listed above by state. Washington, DC, boasts the highest rate, with around 59% of the population participating in the workforce. The northeastern U.S. contains 4 out of the top 15 states in labor force participation.

---

**Question 6**: *What are the 5 states with the lowest average median household income?*

**Relational Algebra**:

```
TOP 5 (τ avg_income DESC (γ State;
AVG(Median_Household_Income_2023) →
avg_income(Employment)))
```

**SQLite Query**:

```
SELECT State,
ROUND(AVG(Med_Household_Income_2023), 2) AS
avg_income
FROM Employment
GROUP BY State
ORDER BY avg_income
LIMIT 5;
```

**Table Output**:

*5 Rows*

| | State | avg_income |
|---|---|---|
| 1 | MS | 47753.02 |
| 2 | AR | 49460.11 |
| 3 | WV | 51542.24 |
| 4 | LA | 52000.28 |
| 5 | AL | 52610.94 |

**Interpretation**:

The five states with the lowest average county-level median household incomes are listed above. Mississippi has the lowest at approximately $47,753. Four of the five states are located in the southern region of the U.S.

---

**Question 7**: *List the number of 2023 births for each Vermont county. Sort the table with the highest number of births at the top.*

**Relational Algebra**:

```
π County, Births_2023 (τ Births_2023 DESC
(σ State = 'VT'(Population)))
```

**SQLite Query**:

```
SELECT County, Births_2023 FROM Population
WHERE State = 'VT'
ORDER BY Births_2023 DESC;
```

**Table Output**:

*14 Rows*

| | County | Births_2023 |
|---|---|---|
| 1 | Chittenden County, VT | 1334 |
| 2 | Franklin County, VT | 524 |
| 3 | Rutland County, VT | 444 |
| 4 | Washington County, VT | 443 |
| 5 | Windsor County, VT | 418 |
| 6 | Windham County, VT | 336 |
| 7 | Addison County, VT | 289 |
| 8 | Bennington County, VT | 263 |
| 9 | Orange County, VT | 248 |
| 10 | Caledonia County, VT | 239 |
| 11 | Lamoille County, VT | 208 |
| 12 | Orleans County, VT | 208 |
| 13 | Grand Isle County, VT | 62 |
| 14 | Essex County, VT | 56 |

**Interpretation**:

This is a table of all 14 Vermont counties and their respective 2023 birth counts. Chittenden County has the highest number because it is also the most populated county in Vermont (it contains Burlington). Essex County is at the bottom of the table, with only 56 births in 2023.

---

**Question 8**: *What is the proportional change in population from 2020 to 2023 for each State? List the states with the greatest proportional change at the top of the table. (Hint: Use the absolute value!)*

**Relational Algebra**:

```
τ ABS(pct_change_pop) DESC
(γ e.State; SUM(Population_Estimate_2023 -
Census_Pop_2020)/ SUM(Population_Estimate_2023) →
pct_change_pop (Population)
```

**SQLite Query**:

```
SELECT State,
       ROUND((SUM(Population_Estimate_2023 -
Census_Pop_2020) * 1.0) /
SUM(Population_Estimate_2023), 4) AS
pct_change_pop
FROM Population
GROUP BY State
ORDER By ABS(pct_change_pop) DESC;
```

**Table Output**:

*50 Rows*

| | State | pct_change_pop |
|---|---|---|
| 1 | ID | 0.0639 |
| 2 | SC | 0.0475 |
| 3 | FL | 0.0474 |
| 4 | TX | 0.0445 |
| 5 | MT | 0.0429 |
| 6 | UT | 0.0428 |
| 7 | DE | 0.0406 |
| 8 | AZ | 0.0377 |
| 9 | NC | 0.0366 |
| 10 | SD | 0.0355 |
| 11 | NY | -0.0322 |
| 12 | TN | 0.0303 |
| 13 | GA | 0.0288 |
| 14 | NV | 0.028 |
| 15 | ME | 0.0239 |

**Interpretation**:

Above are the proportional changes in population from the 2020 Census to the 2023 estimated population by state and sorted by magnitude of change. States with the largest change in population are at the top of the table. The states with the largest change have almost all been increases, with Idaho, South Carolina, Florida, and Texas having the largest increases. This could be explained by retirees migrating to southern states or immigrants entering states along the U.S. border. New York had the largest proportional decrease in population since 2020. The COVID-19 pandemic could explain this, as it drove many residents out of New York City and into other states.

---

**Question 9**: *Which rural/urban code has the highest average median household income in 2023?*

**Relational Algebra**:

```
TOP 1(τ avg_income DESC
(γ Rural_Urban_Code_2023; AVG(Med_Household_Income_2023) → avg_income
(Poverty ⋈ Poverty.County = Employment.County Employment)))
```

**SQLite Query**:

```
SELECT p.Rural_Urban_Code_2023,
     ROUND(AVG(Med_Household_Income_2023), 2) AS avg_income
FROM Poverty p
JOIN Employment e
ON p.County = e.County
GROUP BY Rural_Urban_Code_2023
ORDER BY avg_income DESC
LIMIT 1;
```

**Table Output**:
*1 Row*

| | Rural_Urban_Code_2023 | avg_income |
|---|---|---|
| 1 | 1 | 82763.14 |

**Interpretation**:

As mentioned above, rural-urban codes classify U.S. counties based on metro status and population. The table above shows that counties with a rural-urban code of 1 (metro counties with at least 1,000,000 residents) have the highest average median household income among all codes, with annual earnings of approximately $82,763. This result makes the most sense because largely populated metro areas are some of the most expensive areas in the United States.

---

**Question 10**: *What are the top ten counties with the lowest unemployment rates in 2023 among counties with an estimated 2023 population of over 400,000?*

**Relational Algebra**:

```
TOP 10(τ
Employment.Unemployment_Rate_2023 DESC(π
Population.County,
Employment.Unemployment_Rate_2023(σ
Population_Estimate_2023 >
400000(Employment ⋈ Employment.County =
Population.County Population))))
```

| | County | Unemployment_Rate_2023 |
|---|---|---|
| 1 | Miami-Dade County, FL | 1.8 |
| 2 | Anne Arundel County, MD | 1.8 |
| 3 | Montgomery County, MD | 1.9 |
| 4 | Madison County, AL | 2 |
| 5 | Baltimore County, MD | 2.2 |
| 6 | Prince George's County, MD | 2.2 |
| 7 | Hillsborough County, NH | 2.3 |
| 8 | Dane County, WI | 2.3 |
| 9 | Johnson County, KS | 2.4 |
| 10 | Charleston County, SC | 2.4 |

**SQLite Query**:

```
SELECT p.County,e.Unemployment_Rate_2023
FROM Employment e JOIN Population p ON e.County = p.County
WHERE p.Population_Estimate_2023 > 400000
ORDER BY e.Unemployment_Rate_2023 LIMIT 10;
```

**Table Output**:
*10 Rows*

**Interpretation**:

Among counties with a population over 400,000 people, Miami-Dade County in Florida has the lowest unemployment rate of 1.8%. Four out of the top ten counties are located in Maryland. This can be attributed to their close distance to Washington, DC, and Baltimore. Also listed at number four on the list is Madison County, Alabama. Considering this state is the only state on the list in the South, it appears as an outlier. Madison County is a hub for space-related activity by NASA, which largely contributes to the low unemployment rate.