ETL Report – Mastery Project

Darrell Gerber  
December 14, 2021

# Introduction

The Mastery project asks us to learn more about the quality of life in the US. The two main questions are 1) What are the factors that impact the quality of life, and 2) Could the factors of quality of life vary from area to area. Two required data sets were provided with the option of using other data sets as needed to support our hypothesis.

# Data Sources

## Required Datasets:

**Behavioral Risk Factor Data: Health-Related Quality of Life**

Centers for Disease Control. Behavioral Risk Factor HRQOL (1993-2010) [Data file]. Retrieved from https://data.world/cdc/behavioral-risk-factor-hrqol

Accessed December 13, 2021.

Summary - 1993 - 2010. Centers for Disease Control and Prevention (CDC). Data are from the Behavioral Risk Factor Surveillance System (BRFSS). All respondents to the BRFSS are non-institutionalized adults, 18 years old or older. HRQOL surveillance is used to identify unmet population health needs including recognizing trends, disparities, and determinants of health in the population. HRQOL surveillance data can be used to inform decision making, and program and policy development. To assure that the population is benefiting from public health programs, HRQOL surveillance data can be used for program evaluation. A compact set of HRQOL measures including a summary measure of unhealthy days have been developed and validated for population health surveillance and have been widely used since 1993.

**US Population by Zip Code**

US Census Bureau. (2000 and 2010). US Population By Zip Code (Version 1) [Data file]. Retrieved from https://www.kaggle.com/census/us-population-by-zip-code

Accessed December 13, 2021.

Summary: The United States census count (also known as the Decennial Census of Population and Housing) is a count of every resident of the US. The census occurs every 10 years and is conducted by the United States Census Bureau. Census data is publicly available through the census website, but much of the data is available in summarized data and graphs. The raw data is often difficult to obtain, is typically divided by region, and it must be processed and combined to provide information about the nation as a whole.

The United States census dataset includes nationwide population counts from the 2000 and 2010 censuses. Data is broken out by gender, age and location using zip code tabular areas (ZCTAs) and GEOIDs. ZCTAs are generalized representations of zip codes, and often, though not always, are the same as the zip code for an area. GEOIDs are numeric codes that uniquely identify all administrative, legal, and statistical geographic areas for which the Census Bureau tabulates data. GEOIDs are useful for correlating census data with other censuses and surveys.

Where did you find your data? When did you access it? Cite your sources here. Use proper citation here, using the APA format.

## Additional Datasets:

**Zip Code by State**

National Bureau of Economic Research. Individual Income Tax Statistics - ZIP Code Data (SOI) (2017-06-21) [Data file]. Retrieved years 2005 and 2010 from https://www.kaggle.com/census/us-population-by-zip-code

Summary: The IRS creates Individual Income Tax ZIP Code Data which contains selected income and tax items classified by State, ZIP Code, and size of adjusted gross income.

# Extraction

All of the files were downloaded as CSV files and imported into Power Query in Excel as Unicode (UTF-8) for further transformation.

# Transformation

The five datasets mentioned above will be combined to three datasets using Power Query in Excel. Behavioral\_Risk\_Factor\_Data consists of the contents of the Behavioral Risk Factor Data: Health-Related Quality of Life dataset with columns removed. PopByState2010 consists of a combination of the Population by Zipcode for 2010 and Zip Code by State for 2010. PopByState2000 consists of a combination of the Population by Zipcode for 2000 and Zip Code by State for 2005. Note: 2005 was used as the nearest available date to the Population by Zipcode date.

**Behavioral\_Risk\_Factor\_Data**

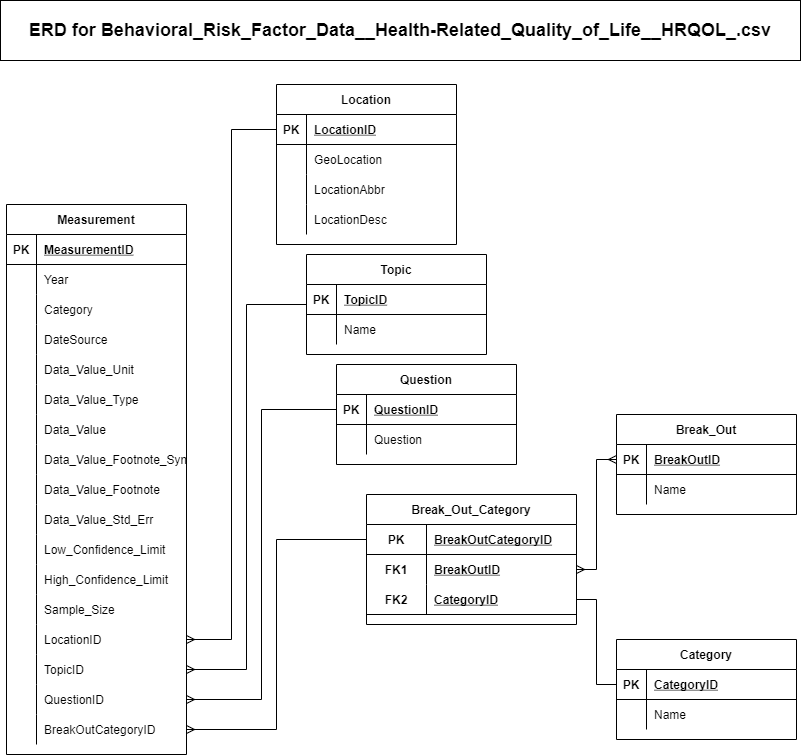
|  |  |
| --- | --- |
| **File** | **Transformation** |
| Behavioral\_Risk\_Factor\_Data\_\_Health-Related\_Quality\_of\_Life\_\_HRQOL\_.csv | Removed Column DataSource |
|  | Removed Column Data\_Value\_Unit |
|  | Removed Column Data\_Value\_Std\_Err |
|  | Removed Column CategoryID |
|  | Removed Column LocationID |
|  | Removed Column TopicID |
|  | Removed Column BreakoutID |
|  | Removed Column BreakoutCatID |
|  | Removed Column Data\_Value\_Footnote\_Symbol |
|  | Removed Column Data\_Value\_Footnote |
|  | Removed Column Category |

**PopByState2010**

|  |  |
| --- | --- |
| **File** | **Transformation** |
| population\_by\_zip\_2010.csv | Add Column ‘zipcode\_text’ the converts the zipcode to text adding the leading zeros. Number.ToText([zipcode], "00000") |
|  | Change Type of ‘population’ to Whole Number |
|  | Filter ‘gender’ by <blank> to show only the total population for that zip code |
|  | Removed Column ‘gender’, ‘minimum\_age’, ‘maximum\_age’, ‘zipcode’, and ‘geo\_id’. |
| zipcode2010.csv | Add Column ‘zipcode\_text’ the converts the zipcode to text adding the leading zeros. Number.ToText([zipcode], "00000") |
|  | Remove Duplicates in ‘zipcode\_text’ |
|  | Removed Columns – All IRS return data |
|  | Removed Column ‘statefips’ |
|  | Removed Column ‘zipcode’ |
|  | Removed Column ‘year’ |
| PopByState2010 | Merge zipcode2010.csv with population\_by\_zip\_2010.csv with a Left Join on the ‘zipcode\_text’ columns. |
|  | Expand ‘population\_by\_zip\_2010’ table to include only population. Creates a ‘population\_by\_zip\_2010.population’ column. |
|  | Group By ‘state’ to new column ‘StateSum’, operation is Sum on column ‘population\_by\_zip\_2010.population’. This creates the total population by state by adding the population for all zipcodes in each state. |

**PopByState2000**

|  |  |
| --- | --- |
| **File** | **Transformation** |
| population\_by\_zip\_2000.csv | Change Type of ‘population’ to Whole Number |
|  | Filter ‘gender’ by <blank> to show only the total population for that zip code |
|  | Removed Column ‘gender’, ‘minimum\_age’, ‘maximum\_age’, and ‘geo\_id’. |
| zipcode2005.csv | Add Column ‘zipcode\_text’ the converts the zipcode to text adding the leading zeros. Number.ToText([zipcode], "00000") |
|  | Remove Duplicates in ‘zipcode\_text’ |
|  | Removed Columns – All IRS return data |
|  | Removed Column ‘zipcode’ |
| PopByState2000 | Merge zipcode2005.csv with population\_by\_zip\_2000.csv with a Left Join on the ‘zipcode’ and ‘zipcode\_text’ columns. |
|  | Expand ‘population\_by\_zip\_2000’ table to include only population. Creates a ‘population\_by\_zip\_2000.population’ column. |
|  | Group By ‘state’ to new column ‘StateSum’, operation is Sum on column ‘population\_by\_zip\_2010.population’. This creates the total population by state by adding the population for all zipcodes in each state. |



# Load

Follow these pseudo-code steps to load the Behavioral\_Risk\_Factor\_Data\_\_Health-Related\_Quality\_of\_Life\_\_HRQOL\_.csv into a SQL relational database constructed with the data tables in the diagram above.

1. Load the CSV file
2. Find unique Locationids in the CSV file.
3. Configure to allow setting IDs
4. Load all LocationID, GeoLocation, LocationAbbr, and LocationDesc from the file into the Location table.
5. Find unique Topicids in the CSV file.
6. Load all Topicid, and Topic from the file into the Topic table with Topic going to ‘Name’ in the table.
7. Find unique Questionids in the CSV file.
8. Load all QuestionID, and Question from the file into the Question table.
9. Find unique BreakOutids in the CSV file.
10. Load all BreakOutId, and Break\_Out from the file into the Break\_Out table with Break\_Out going into the Name field in the table.
11. Find unique CategoryIDs in the CSV file.
12. Load all CategoryID, and Category from the file into the Category table with Category going into the Name field in the table. Note, there is only one Category in this file.
13. Find unique BreakOutCategoryIDs in the CSV file.
14. Load all BreakOutCategoryID, BreakOutID, and CategoryID into the Break\_Out\_Category table.
15. Configure to disallow setting IDs
16. Sort to show all measurements.
17. Load Year, DataSource, Data\_Value\_Unit, Data\_Value\_Type, DataValue, Data\_Value\_Footnote\_Symbol, Data\_Value\_Footnote, Data\_Value\_Std\_Err, Low\_Confidence\_Limit, Hight\_Confidence\_Limit, Sample\_Size, LocationID, Topicid, Questionid, and BreakOutCategoryID into the Measurement table.

# Conclusion

The data in the Behavioral\_Risk\_Factor\_Data\_\_Health-Related\_Quality\_of\_Life\_\_HRQOL\_.csv is now loaded into Excel in three data tables. The data in the Behavioral\_Risk\_Factor\_Data worksheet may require additional manipulation via PivotTables to facilitate further analysis, but that will be dependent on the graphic and analysis need and left to later steps. The PopByState2010 and PopByState2000 worksheets are now configured for direct use in analysis and display.