ETL Report Guide

Jassleen Bhullar

December 1, 2022

Introduction

I am creating a report that illustrates factors that affect quality of life and how they vary from area to area. To do this, I must first extract, transform, and load data so that all my data is in one place, and it can easily be queried. To create visualizations that summarize the data, I need the data to be all loaded into one database, and the data must be normalized so that I can reference and query records from different tables at once.

Data Sources

Bureau of Economic Analysis. (2011, September). State Personal Income: Revised estimates for 2010, Version 2. Retrieved December 1, 2022 from https://apps.bea.gov/regional/histdata/releases/0911spi/index.cfm

Centers for Disease Control. (2016, January). Behavioral Risk Factor Data: Health-Related Quality of Life (HRQQL), Version 1. Retrieved November 28, 2022 from https://data.world/cdc/behavioral-risk-factor-hrqol

DC Data Journalism. (2017, January). US Health and Demographic Data: Race_ethnicity.csv, Version 1. Retrieved December 1, 2022 from https://data.world/dc-data-journalism/urban-rural-health-and-demographic-data/workspace/file?filename=Race_ethnicity.csv

Michael Valcic. (2017, January). US Population By Zip Code: Add city, state, longitude, and latitude data, Version 1. Retrieved November 28, 2022 from https://www.kaggle.com/code/mvalcic/add-city-state-longitude-and-latitude-data/script?scriptVersionId=2190548

Extraction

All the datasets I extracted were in CSV format. I clicked the download button that was available on every website, and then the CSV file was in my downloads.

1. Open a Excel workbook.
2. Go to the Data ribbon. Select Get Data. Choose From Text/CSV file.
3. When the preview loads, choose Transform. The file is now in Power Query Editor.

Once the file is open in Power Query Editor, I can start the transformation process.

Transformation

The datasets I used were largely cleaned already, but I still had to do a few transformation steps for each dataset.

CDC Behavioral Risks:

1. Delete columns that seemed irrelevant such as confidence limits, category, categoryID, etc.
2. Added conditional column to separate average data values from percentages.
3. Added new columns that separated the text between delimiters from Geolocation column to separate Longitude and Latitude.
4. Filtered out PR and US from state column.
5. Renamed columns to have better titles.
6. Deleted any other columns I didn't need after further data exploration.

Population by Zip Codes:

1. Removed columns like column 1 and geo_id
2. Filtered out PR
3. Changed data type to number for certain the age and population columns
4. Filtered gender column so I only had blanks
5. Grouped by State, summed population
6. Merged 2000 and 2010 into one query

Income

1. Deleted all columns before 1993
2. Changed data type to number
3. Added a column for state abbreviations
4. Filtered out PR, US, Regions
5. Unpivoted Columns
6. Renamed Columns

Race/Ethnicity:

1. Removed unnecessary columns like state_fips
2. Changed data type of population column to number
3. Filtered out 2015 data
4. Grouped by race and state, summed population
5. Renamed columns
6. Changed race group names so they matched up with risk factors dataset

Load

There are a few steps that must be taken to load transformed data into a SQL database.

1. Close & Load on Power Query
2. Save File as Workbook so that you can return to your Query.
3. Save File as CSV (UTF-8)
4. Close the CSV file.
5. Go to Azure & Find the database you want to import data into from the menu on the right. Make sure you are connected to your SQL Server.

6. Click on Import Wizard. If Import Wizard is not installed, install it as an extension.
7. Specify the Server and Database you want the data in. Specify the location of the file and give the new table you are creating with your data a reasonable name.
8. Click Next. This should load a preview of your data. Make sure it is the data you want to import.
9. Click Next. Allow All Nulls. Make sure data types are accurate.
10. Click Import. Your data should now successfully be imported.

Conclusion

This report details all the ETL I used to create my qol database that I then used to make my report. It should serve as a useful guide to those seeking to recreate my ETL.