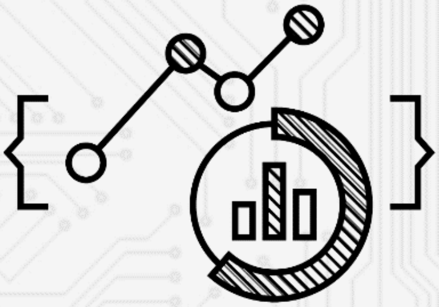


# ETL: Extract, Transform, Load

The goal of this project, as stated in the proposal, was to collect data from various sources on uninsured people in the United States. And in so doing, to answer the question about who these people are. And where they are located. Additionally, I was looking for information on why they had lost or did not have insurance, this proved to be the most difficult.

## ANNUAL REPORT

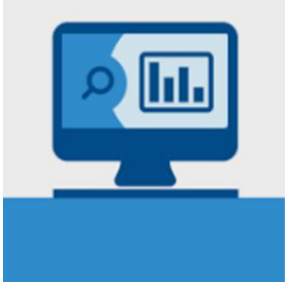
The first data source used was https://www.americashealthrankings.org/Annual Report. The Annual Report is the longest running annual assessment of the nation's health on a state-by-state basis. While this only a numerical measure of sates against a national value, it is useful at comparing quality of care received between states. While somewhat tangential I determined that quality and availability of care was as important as being able to pay for it.

This dataset required little cleaning and was largely unchanged as it was in a simple .CSV format. The two problems with this dataset are that the 2019 report is based on data gathered from various time periods from 2016-2019. And that it is based on a formula that is a little abstract.

## US CENSUS

The second data source was the US census, the table used was "Selected Characteristics of the Uninsured in the United States" This was a large dataset that covered national demographics of the total population as well as data on the demographics of the uninsured population. This was both the most useful and most difficult to work with data set, the .CSV file was optimized to work with a web application and had nonstandard naming conventions. So, while the while was in a .CSV format it proved to be the most difficult to work with.

In addition, each year was a separate table, I pulled data from 2018 to 2010. Each these had to be combined into a single file, I then renamed / cleaned the column names to make that data more readable. This large dataset was the bulk of the work, but has provided the majority of the data, going back eight years. This data could be further cleaned by removing the margin of error columns, as they make up a little less than half of the data set and are seemed to be exclusively in the 0.1-0.2% range.



## DATA CMS

The third data source was Data.cms.gov's percent of the uninsured This was an especially useful dataset as it lays the information out county by county in the US as well as having information about the . I then loaded these into a relational database using PgAgmin4 as a lot of the data, especially in the census data set was only useful when in larger context. Such as the percentage of uninsured in an income bracket was only useful when you know the number of people in that bracket to begin with.