I want to investigate whether impoverished populations have a greater risk of being diagnosed with and dying of cancer. To get an idea of what defines impoverished in this sense, I will look at income, poverty level, and having health insurance. While the demographic data has much more fidelity to it, I will only be able to get down to the county level due to the nature of the cancer data I am using. Often, that is a small enough area to see trends in poverty and income levels, as evidenced by public school performance. I believe this will be acceptable for this investigation. More thorough research should try to get fidelity down to a zip code or neighborhood. I lack the expertise or experience to find it.

This data was compiled from two sources. To get cancer mortality and incidence data I downloaded state-by-state data at the county level from [statecancerprofiles.cancer.gov](https://statecancerprofiles.cancer.gov), which is compiled from many sources of research by the CDC and NIH. The incidence data came from [/incidencerates/index.php](https://statecancerprofiles.cancer.gov/incidencerates/index.php) and mortality from [/deathrates/index.php](<https://statecancerprofiles.cancer.gov/deathrates/index.php>).

I always find the US Census data difficult to get to. Looks like the latest American Community Survey (ACS) data available is for 2015 (2016 will be released later this month). Fortunately, I found the 2015 data readily available on [data.world](https://data.world/), so the demographic data came from three of their datasets: [/uscensusbureau/acs-2015-5-e-poverty](https://data.world/uscensusbureau/acs-2015-5-e-poverty), [/uscensusbureau/acs-2015-5-e-income](https://data.world/uscensusbureau/acs-2015-5-e-income), and [/uscensusbureau/acs-2015-5-e-healthinsurance](https://data.world/uscensusbureau/acs-2015-5-e-healthinsurance).

For the cancer data I had to download each state’s individual cancer profile information. For the demographic data, I downloaded the census data's poverty, income, and health insurance datasets, which are also stored in individual state CSVs. I read the CSVs into Power Query as a folder to combine and transform them.

For the cancer data, I used Power Query to strip out all of the blank rows and all of the table detail and table notes included in the CSV downloads. I padded the FIPS fields with leading zeros to ensure all were 5 digits, and renamed the fields for simplicity and to ensure they were legal for R. I had to make a decision because both sets of cancer data had counties that reported no data stating there were three or less incidences/deaths over the last 5 years. This seemed unlikely for some, but I did not have time to investigate further much less research data to backfill where it could be. I decided the loss of 260 counties worth of data out of 3,220 was acceptable.

For the demographic data, I filtered the data to summary level 50 to only bring in the county-level data to match the incidence and mortality datasets. I then used the column key files to find the fields I was interested in investigating then remove the superfluous fields. Utilizing Power Query further, I combined the state and county FIPS fields into a single FIPS field to also make it compatible with the incidence and mortality datasets. I padded the FIPS field with leading zeros to ensure all were 5 digits, and renamed the fields for simplicity and to ensure they were legal for R. Finally, I used Merge Queries as New to combine all the CSVs into a single CSV file to upload to the server.