|  |
| --- |
| Final Project  Data Investigation for Cancer Trials |
| Bruce GoldfederCSI 773December 13, 2016 |

# Abstract

I am extending the work on the Data.World Clinical Trials [1] open repository for the analysis and investigation of Clinical Trials for Cancer. The owner of the project, Noah Rippner has added me as a contributor to this project. This is the Data Science equivalent of a software open source project where members can join to investigate a common interest. This involves the open sharing of data sources, data manipulation scripts, methodologies, and various models for investigating the data. I will be contributing back to the project the final project materials to include all code, plots, and the text of this final project. The data set is put together using economic statistics, clinical trial statistics, and geospatial data to examine the relationship between cancer research, economic status, and location within the United States.

# Introduction

For the redesign project, I analyzed a time series data set involving Cancer Incidence Rates between 2009 and 2013. During this investigation, I encountered more questions than answers and sought out additional data sources to help explain the patterns I was observing in the data. The intent of this project is to apply the statistical discovery methods learned in our course to further investigate the data set and facilitate discovery of models and any resultant patterns by adding dimensions for clinical trials, cancer rates, socioeconomic status, death rates, and geospatial location.

# Report

## Data Sets

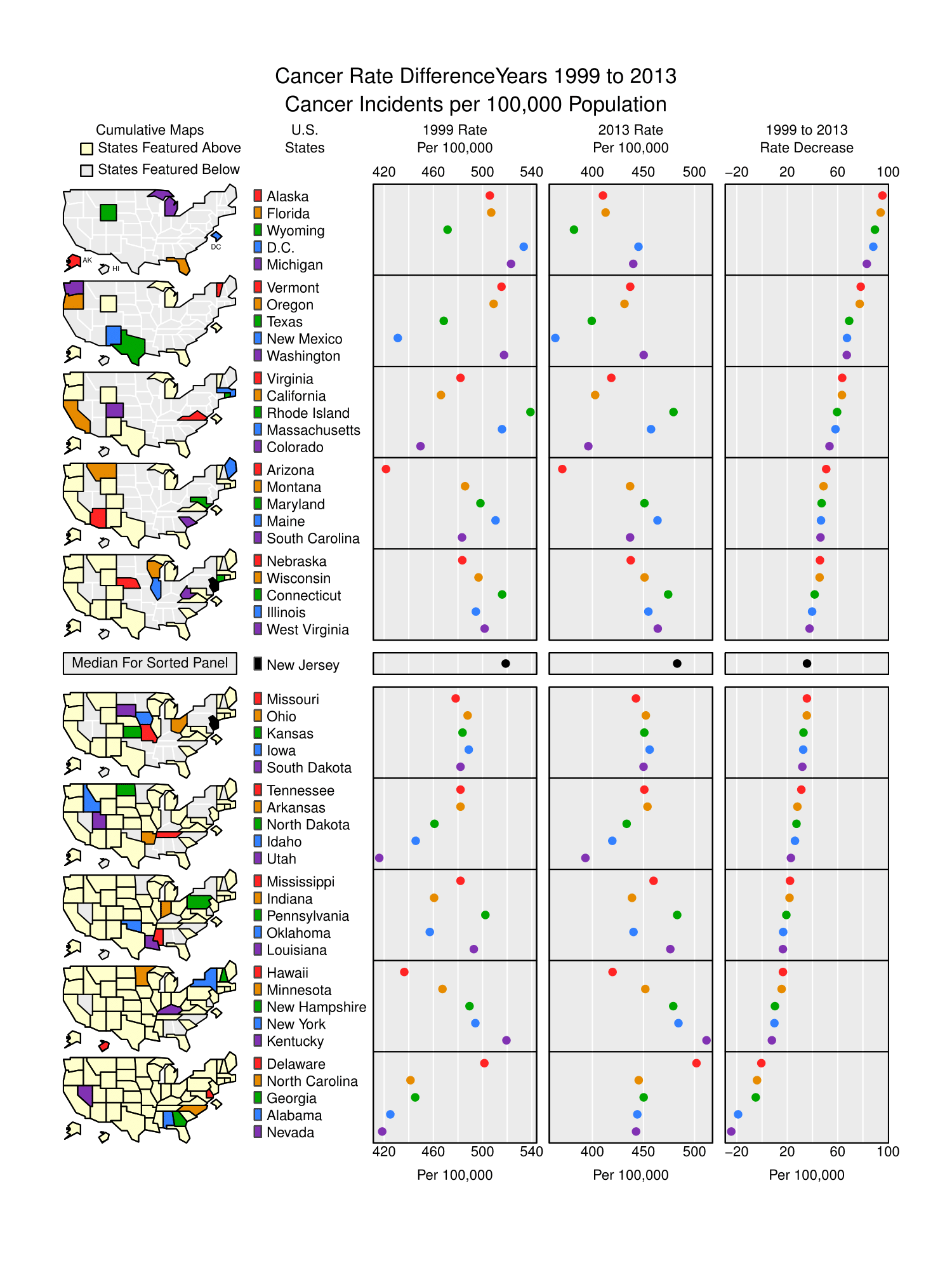
The data set provided by the Cancer Trials project combine multiple data extracts from three separate sources, research information from clinicaltrials.gov [2], incidence of cancer and death rates from cancer.gov [3], and economic and population statistics from census.gov [4]. In aggregate the raw data sets comprised of millions of records with over one hundred columns of data. The data set used in this project are filtered (row and column) and merged as required to fit various analytic techniques and visualization tools.

## Analytic Tools

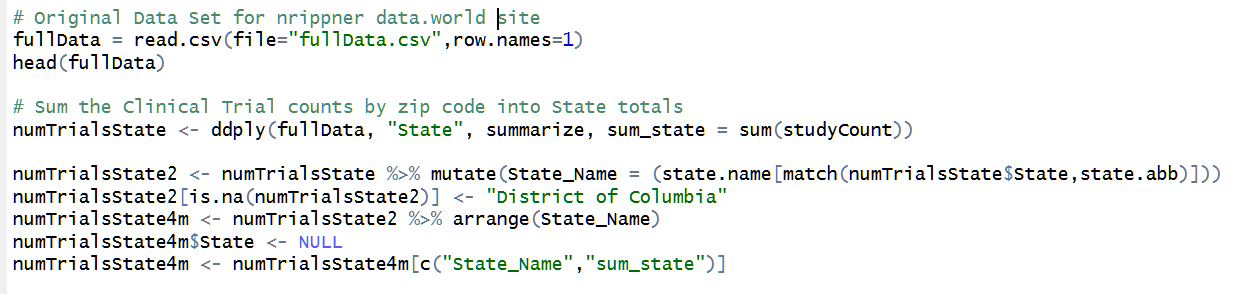
The principal tool for analytics will be R with scripts created in R-Studio and leveraging the packages learned and exercised in class. All the graphs and supporting code were created with the tools learned in class augmented with occasional assists from StackOverflow [5] for coding suggestions and analytic interpretation assistance.

# Data Investigation

The starting point for this investigation is the time-based analysis of Cancer Incident rates over the period from 2009 through 2013 (recreated with your recommendations).

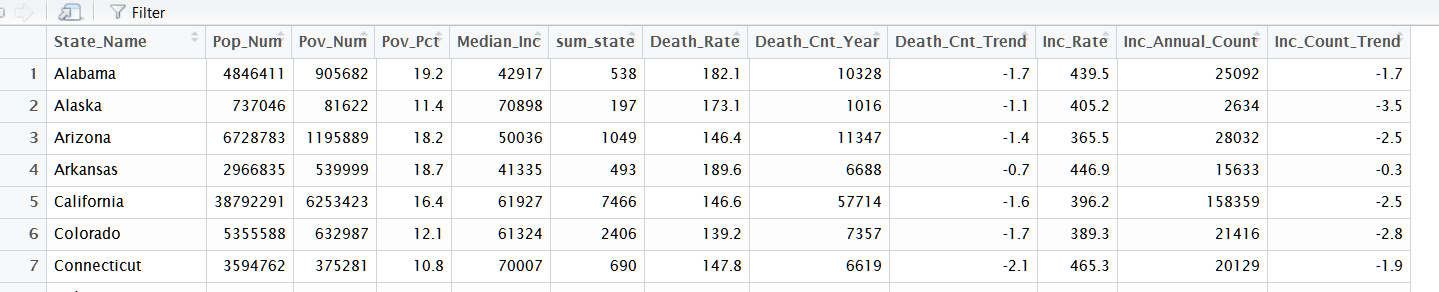


To investigate the reasons for the odd patterns and the fact that Alaska had the largest drop in Cancer Incidence, I sought additional dimensions such as socio-economic data, population levels, affluence of the state, and the prevalence of clinical trials. These added dimensions were downloaded from four separate queries into the repositories described and cited previously. Extensive use of the R library “dplyr” as taught in class made filtering and merging the four data sets much easier. A snippet of the several pages of code is shown below:



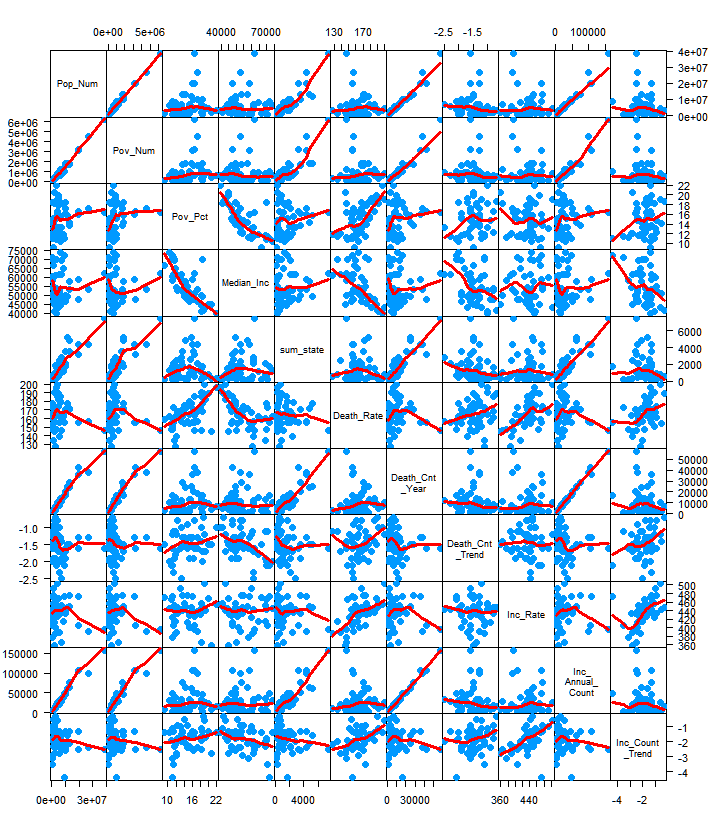
Even with assistive functions of dplyr, there were missing data, bad data, character/number issues, and the R state.name function not having District of Columbia in it.

I always start my investigations into data by looking at the raw data. Below is the data set I created.



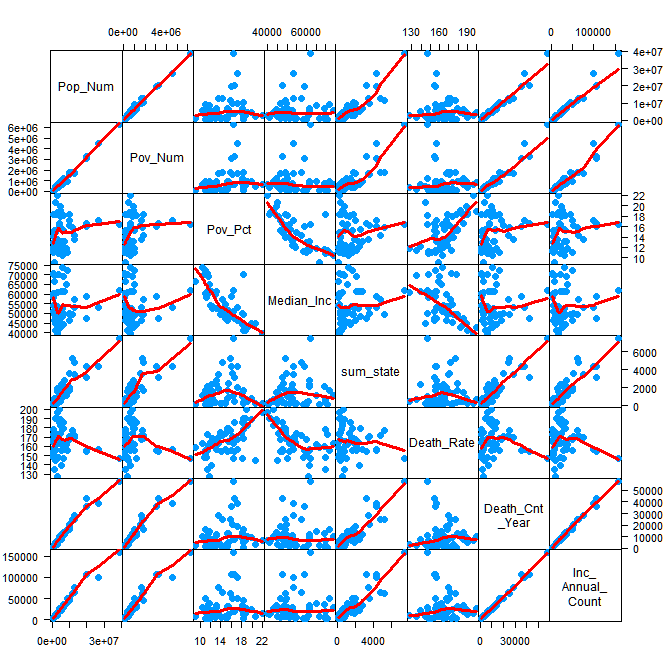
Much of the data had to be normalized to the State level as the original investigation had a majority of the data at the county and sub-county level. The clinical data counts were aggregated using R, rather than re-write the python code from the nrippner/Clinical Trials site [1]. That captured counts for over 5300 separate clinical studies. Within R Studio I could sort and filter the data set using the column header sort capability and the filter tool.

## Scatterplot Matrix

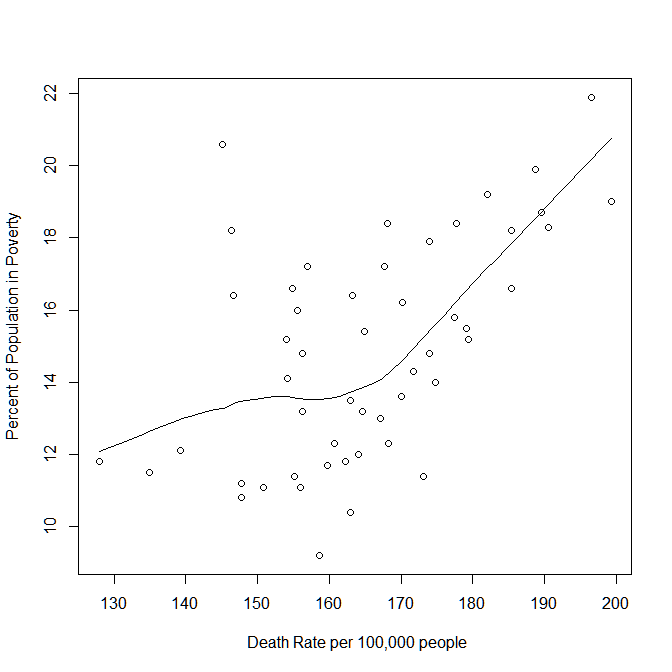
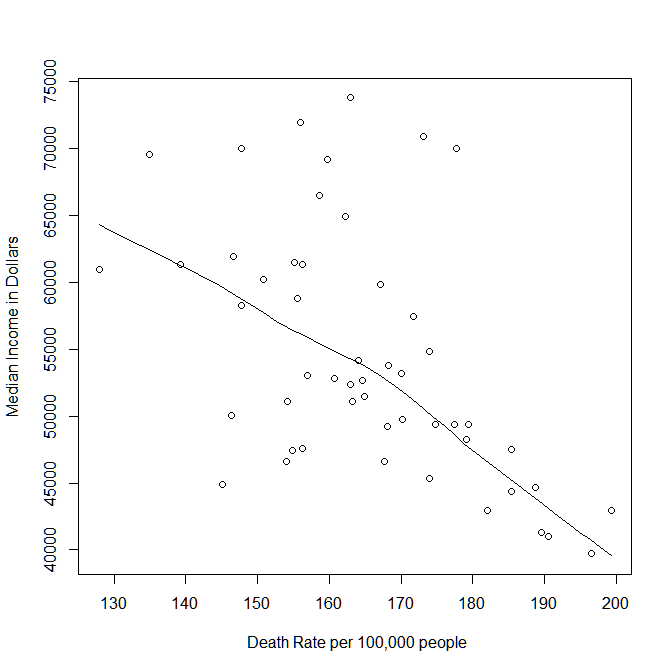
My next step after reading and filtering the data was to create a scatterplot matrix to look for any obvious patterns which depict linear relationships between variables.

Several linear relationships are observable and there are multiple locations where there are outliers that can be investigated. The next step I took was to cull down the number of dimensions to those that I observed had minimal linear interactions. The list to remove were the Incidence Rate (Inc\_Rate), Death Count Trend (Death\_Cnt\_Trend), Incidence Count Trend (Inc\_Count\_Trend). Most of the others had at least one or more striking and obvious linear relationships. The one that immediately jumped out at me were the positive linear correlation between Poverty Percentage (Pov\_Pct) and Death Rate (Death\_Rate). This is juxtaposed with the negative correlation of Median Income (Median\_Inc) with the Death Rate (although there is a leveling off that bears a closer look!)

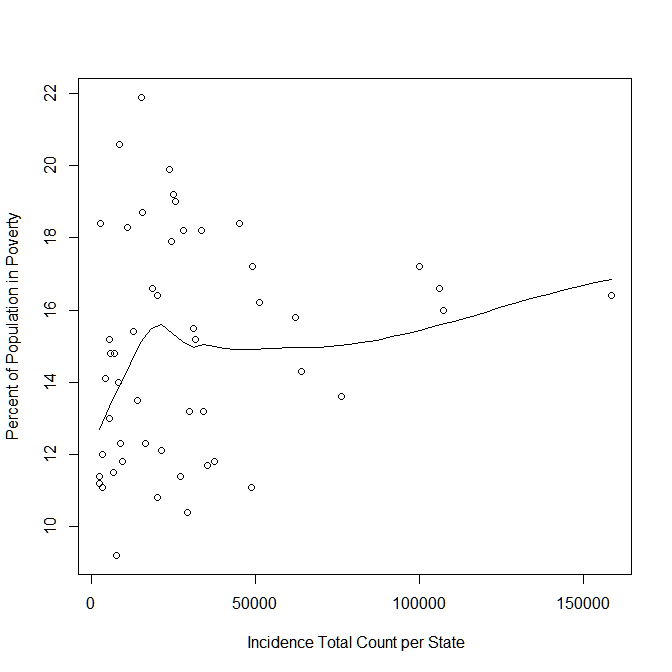
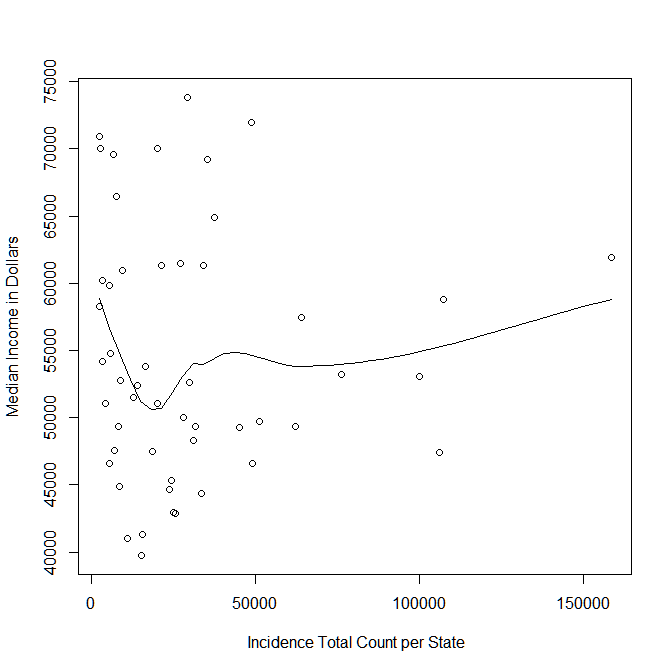
The new scatterplot matrix is more readable and the relationships are mostly obvious, such as higher counts of cancer incidence and cancer deaths where there are higher populations. Population and Poverty are also logically correlated although there are outliers in the states with higher populations. This plot is shown below:



Taking a closer look at the interesting relationship between median income and death rate is made more impactful when juxtaposed with the median income and cancer incidence rate which is neutral.



The following graphs depict the relationships between cancer incidence rates in relation to income and poverty.



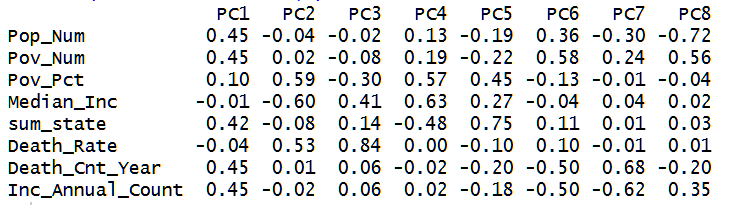
These two pairs of juxtaposed graphs illustrate that although the Incidence of cancer is relatively neutral in relationship to economic status, the death rate is higher for the poor.

## Principal Component Analysis

My next investigation was to perform a Principal Component Analysis (PCA). I began by using the R function prcomp on the model defined previously. The standard deviations of the principal components came out as follows:

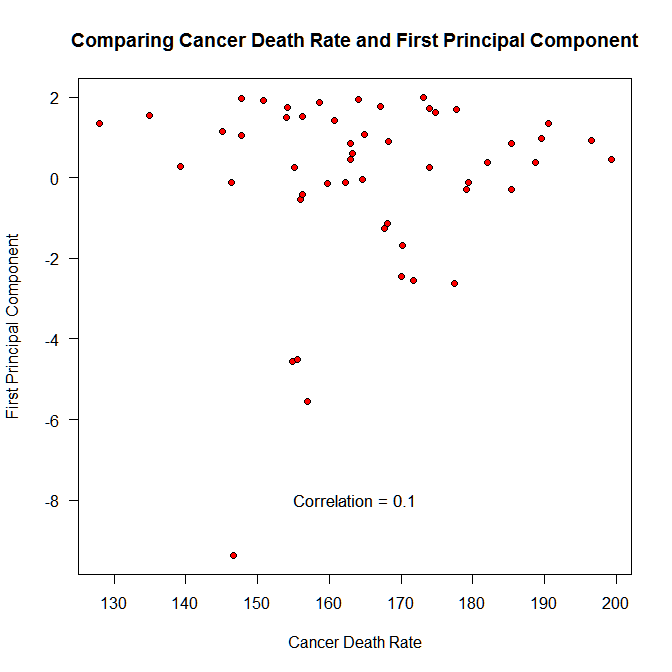
[1] 2.19367534 1.49680685 0.71884611 0.50014667 0.36837617 0.20072150 0.05293793 0.04097351

This did not impact me much, however the rotation matrix had some popouts even in text format:

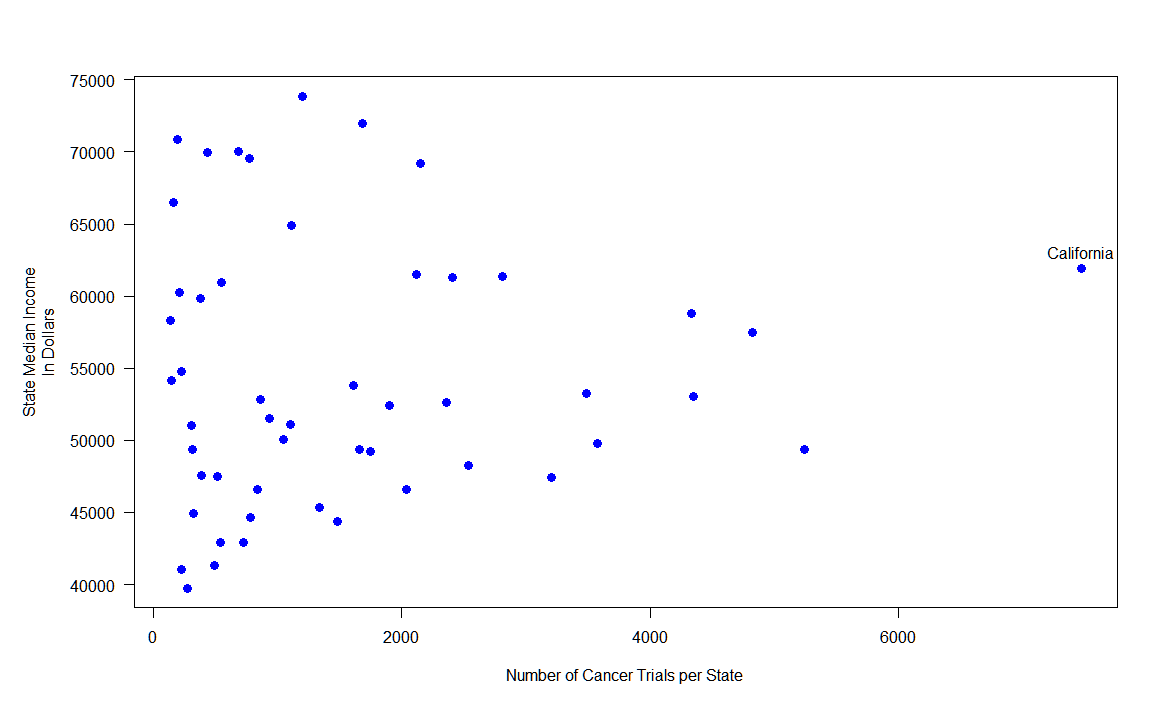


The Death Rate magnitude in the third column sticks out at 0.84 with Median Income and Poverty also having large magnitude entries. This is consistent with our observed graphs so far.

The next examination was to determine if there was a relationship with the Death Rate and the First Principal Component. The graph which is shown on the following page did not have much pattern to it and the calculated correlation shown on the graph was very low at 0.1. I would have expected this to have more of a linear appearance, however those linearities were in contrast to income levels. With respect to the first principal component it is quite neutral.

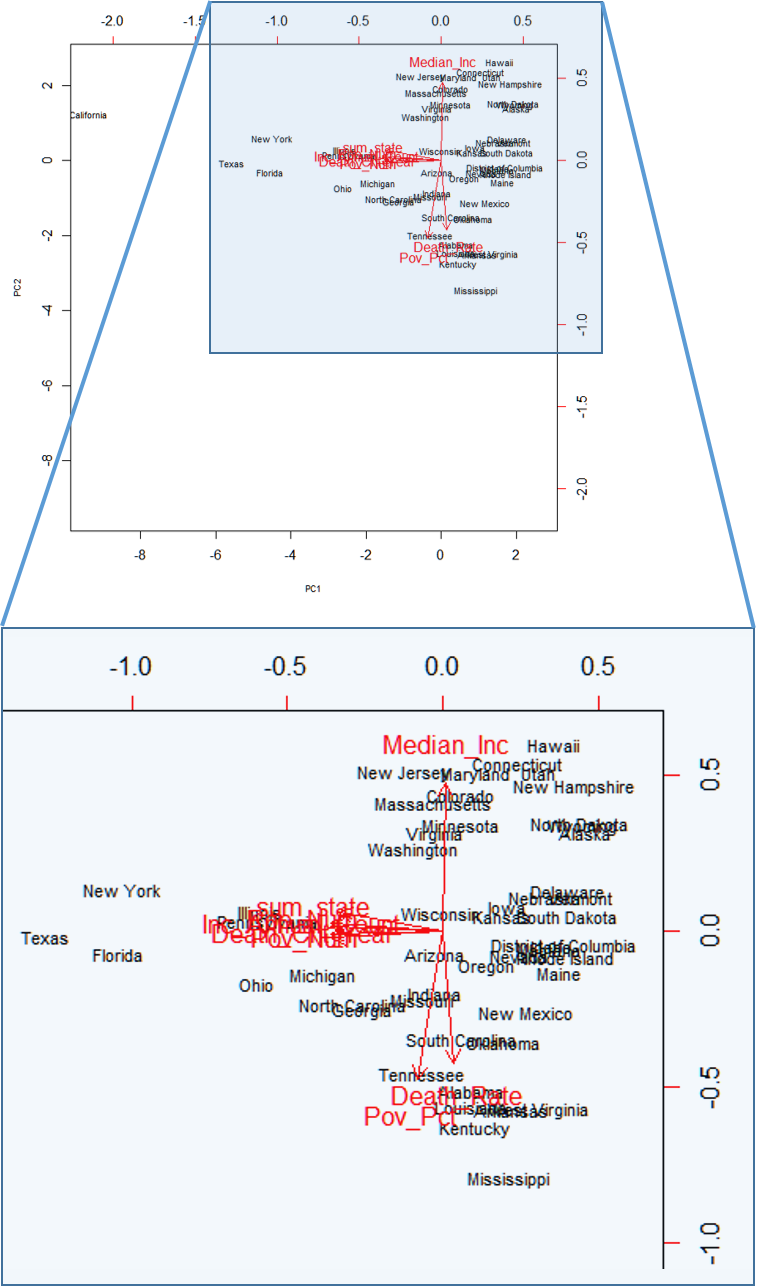


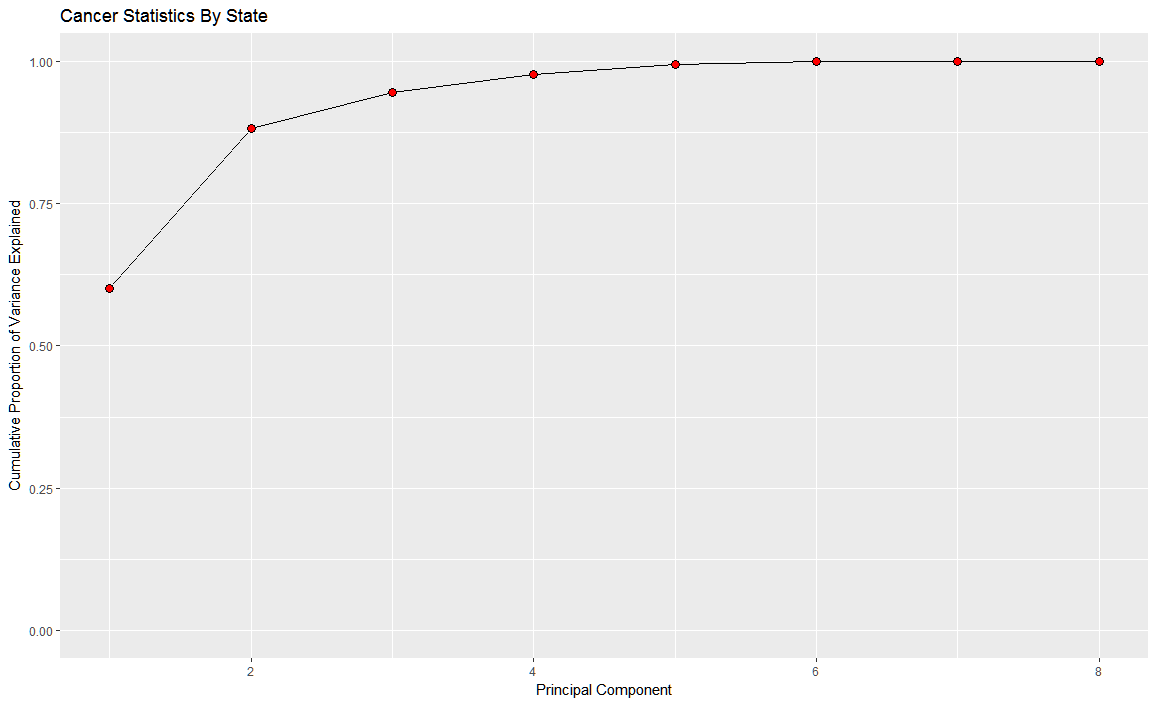
The next examination I pursued were outliers that I noticed from the scatterplot matrix. To review these I leveraged the R function identify. The Summation of the Clinical Trials per State (sum\_state) provided an interesting outlier at the top of each of its plots, I selected two to examine, Median Income and Death Rate.



The identification of California as having the largest number of clinical trials is consistent with it being the state with the highest population.

## Biplot PCA and Percent of Variance graphs

I was still not getting the impact of the PCA investigation, so I decided to explore further by creating a biplot and then creating plots to better explain the percent variance. The biplot is shown below in two parts, the top being the full plot and the bottom a zoom in while dropping the outlier of California, for better readability and analysis of the score of each case (principal components) and the loading of each variable (states). The left and bottom axes are the principal component scores and the top and right axes show the variable loadings. The 2D diagram is a projection of a 3D (or p-space where p=8 in this case) set of orthogonal vectors which is a depiction of the principal components. The first principal axis which maximizes the variance and the second one maximizes the remaining variance. That is the logic of using the first two components to best represent the variable space when it is projected on the plane. As the principal components are comprised of linear combinations or the columns in the model plotting individual factor scores can serve to group states that holistically (all variables) are more homogeneous. Prior to applying this to the cancer statistics p-variables I would like to show the cumulative graph of the percentages of the variances to help explain.



Principal Component variance percentages

[1] 0.6015264352 0.2800538415 0.0645924654 0.0312683360 0.0169626255

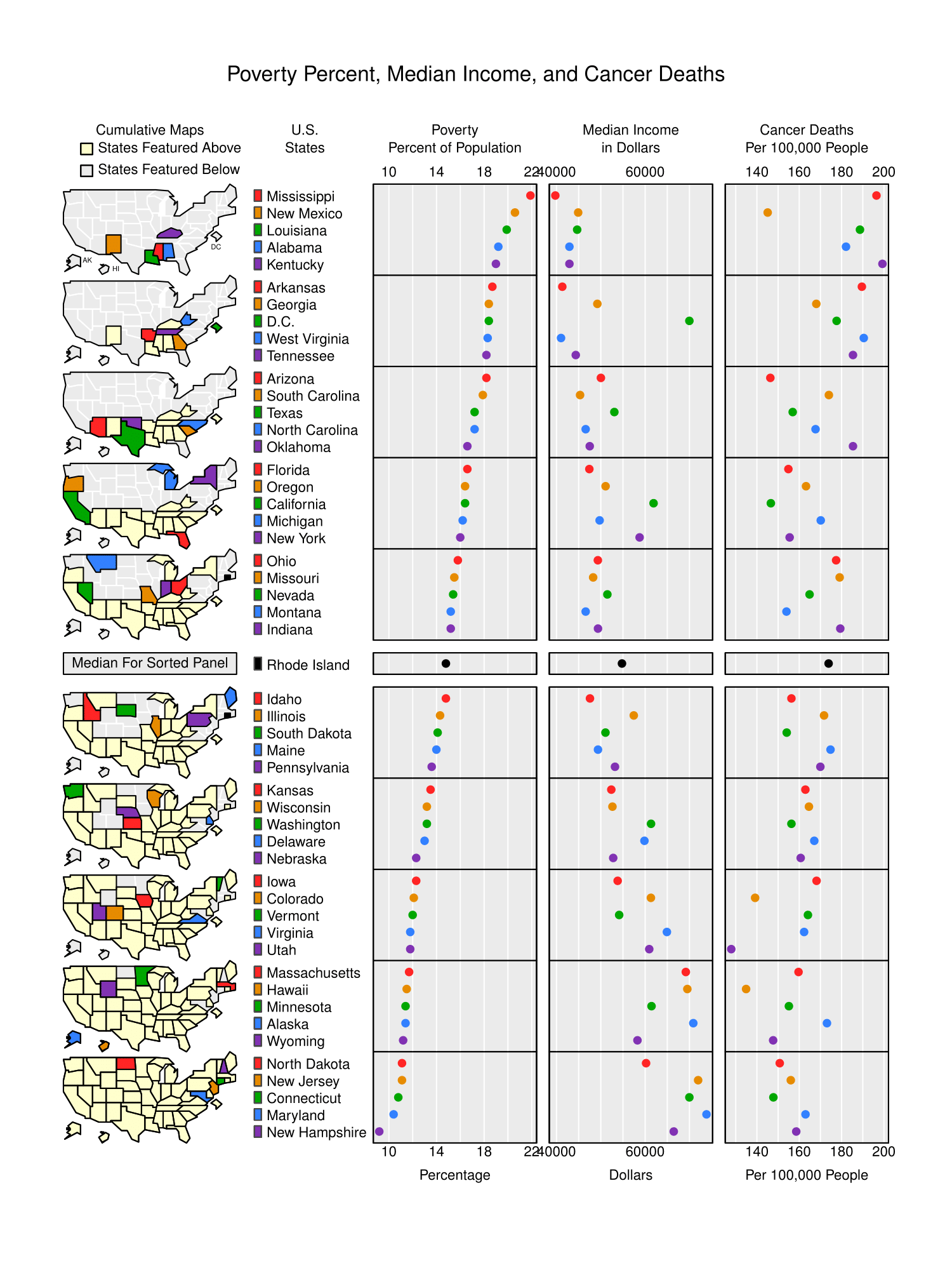
0.0050361399 0.0003503030 0.0002098536

The first principal component comprises just over 60% of the variance. The second brings it up to just over 88% with near 100% reached by the fifth. There are two sets of outliers that are outside of the main grouping which are California with a very strong negative score on the first axis and slightly positive on the second axis. The other grouping is New York, Texas, and Florida which are also negative on the first axis and at or below zero on the second axis. Both groupings score well with the negatively pointing variables (top numbers in most), the variable anomalies in the 2D projection appear as our “usual suspects” while examining this data set, Median Income, Death Rate, and Poverty Percentage. This serves as a validation of our prior explorations which were based on more subjective interpretations of the plots. The states that should perform best at Death Rate and Poverty Percentage (a negative value to optimize) includes Virginia, which is an overall negative homogenous grouping from a health and socio-economic perspective.

## Communication of Exploration Findings

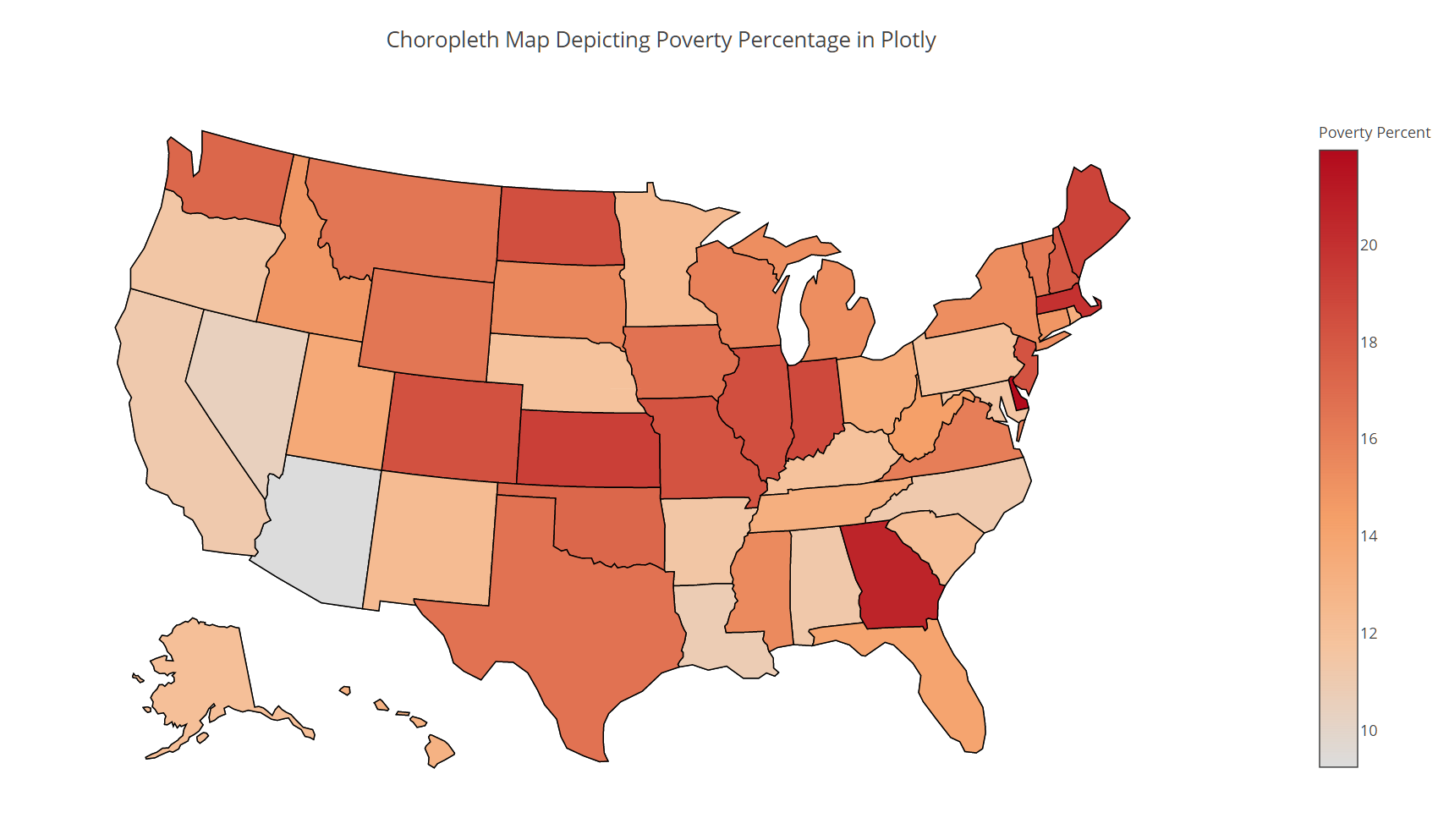
From my exploration of the data set I would now like to incorporate analysis and presentation of the data to begin looking for patterns that can answer some of the questions that arose during my redesign presentation. Based on the scatterplot matrix and further supported by the biplot and PCA analysis I would like to investigate the variables Median Income, Death Rate, and Poverty Percentage as they relate to the rest of the variables. I would also like to incorporate a geospatial representation of the data both to help discover patterns and help communicate the information.

Taking the fruits from the analysis I began where I started by creating a micromapST based visualization using a cumulative plot with three columns representing the three variables that kept being highlighted during this investigation, Poverty Percentage, Median Income, and Death Rate. After a few different sorting combinations, the pattern that first struck me during the scatterplot matrix and was reinforced during the Principal Component Analysis phase became very apparent. The positive linear correlation between poverty and cancer death rates was evident. Also depicted is the inverse, that higher Median Income levels of a state corresponded to lower deaths due to all types of cancers for all people. A possible answer to the question posed during my redesign presentation is evident also in the plots. It seemed odd that Alaska would have the largest drop in cancer incidence rates based on the time series analysis I performed at the time. To support any hypothesis I needed to incorporate additional data dimensions, which I did in this project. Alaska is third in Median Income out of 50 states and ranks seventh lowest in Poverty percentage of the total population. This follows the two patterns. The micromap plot is shown on the next page.

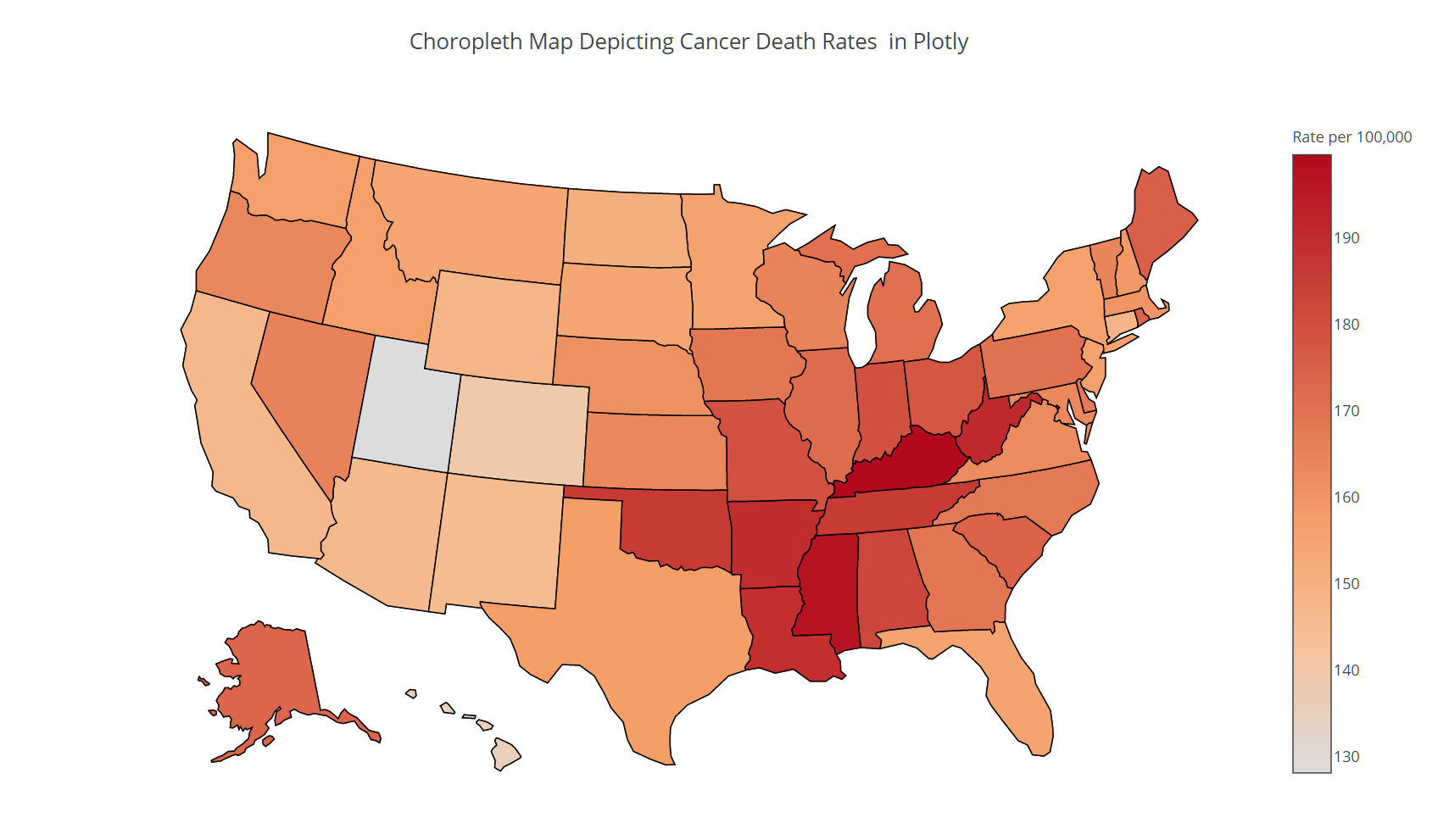


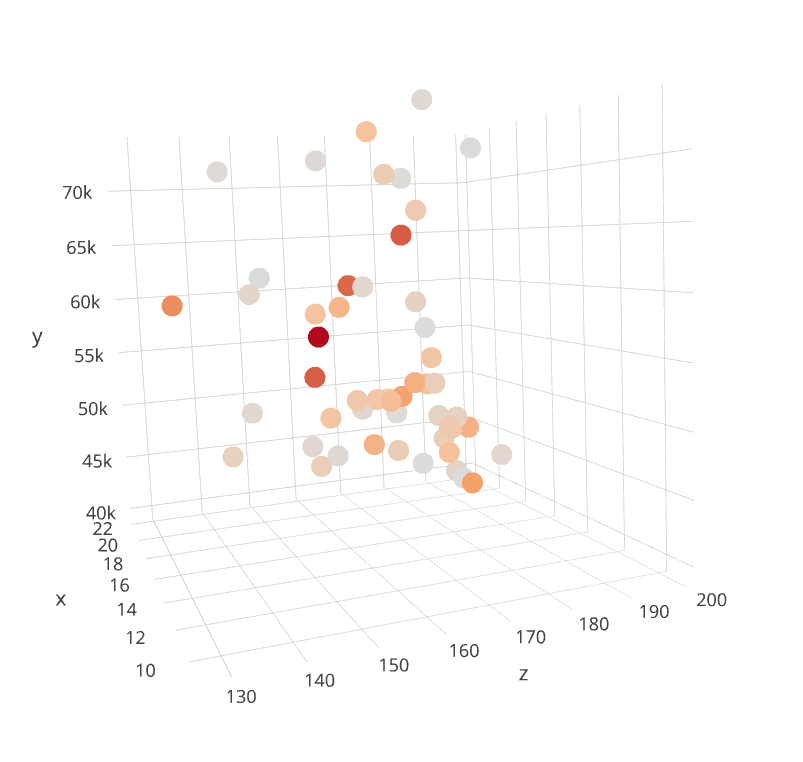
# Discussion and Future Work

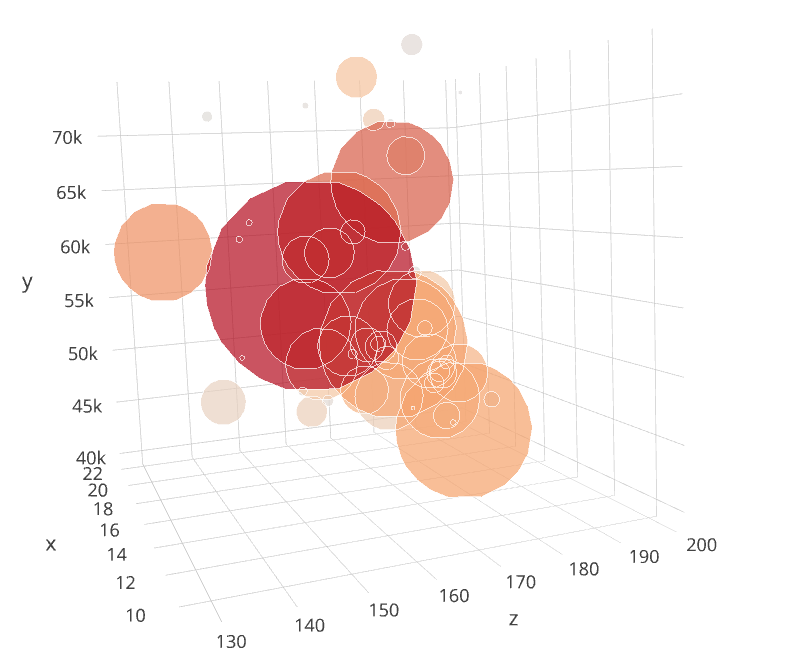
As stated in the introduction, I will be cleaning up and uploading all the R code, supporting graphics, and this report back up to the Data.World/nrippner/DataTrials [1] site. In my discussions with the fellow members of the community there is the desire to delve deeper into the Clinical Trial information to add additional dimensions to the data. There is the hope that in researching the papers additional dimensions may be referenced such as environmental issues, pollutants concentrations, and cultural proclivities to cancer causing behaviors such as smoking. The additional data can be added to the mix and evaluated in the same fashion as well as incorporating alternative methodologies such as random forest to discover new models and patterns within the data. I have also been working with the Shiny R package presented in class and have set up an account on the RStudio sponsored site shinyapps.io. Another tool that I will be leveraging for interactivity is Plotly. I have created Choropleth maps using color as an indicator for Poverty Percentage with mouse over capability to dynamically show the data for individual states. This can also be used with “Bubble” type objects which have both a color and a size component for adding dimensions to lat/lon specified data. I will be demonstrating the capability during my session in class, below is a screen shot of the dynamic tool. An interesting human computer interaction component is that the data is always at the top of the screen and filtering, sorting, editing, and other manipulations can happen in real time during data analysis. The tool is online and is accessible via a web browser but is a rich and intuitive Integrated Development Environment not dissimilar to RStudio.



Although this is a simple depiction, it literally took seconds to get up and running and had the advantage of being able to toggle between different variables very quickly to get a nice visual differencing that was distinctive. In particular, when I toggled to the Death Rate variable I was surprised that the correlation was not greater. During my earlier investigations those two variables showed the strongest correlation. However, it did show me that the data and the variables that I was working with were not showing the full picture. Additional information will be needed to address these discrepancies. The second Choropleth is shown below and is not the inverse of the one above. During the demonstration, I will toggle the Poverty and Income data and the color inverse is very noticeable. The image being on the following is a good example of your cognitive process for tracking changes being reduced when there is a blank space between images. This problem is overcome with the Plotly and R Shiny interactive tools.



Like most Rapid Application Development tools getting to the quick and easy win is simple. For example, I was able to create a 3D depiction of the three variables and have a visualization to investigate which had a nice linear pattern. You could add additional dimensions using color and size. Care needs to be taken to not overload with too many physical, color, and size dynamics as the complexity can easily overwhelm both a data analyst and end users. This is shown as a screen shots below.



However, to add in complex business rules and create re-usable components a developer will just as rapidly move back to a scripted environment such as R. The limitations will be presented in real time during class, but there is minimal ability to customize and integrate with other tools. Towards this end, the plotly tools are now available as a package in R which I have prototyped. Although I did not have time to create a purposeful graphic for this project, I was able to bring up a quick prototype. Note that the dynamic abilities available on the online version are fully capable within R Studio. Below is a diamond example with hover over identify ability I will show in class.



# Bibliography

|  |  |
| --- | --- |
| [1] | N. Rippner, "Cancer Trials," 28 November 2016. [Online]. Available: https://data.world/nrippner/cancer-trials. |
| [2] | "Cancer Trials," National Institue of Health, 2016. [Online]. Available: https://clinicaltrials.gov. [Accessed 28 November 2016]. |
| [3] | "State Cancer Profiles," National Cancer Institute, 2016. [Online]. Available: https://statecancerprofiles.cancer.gov/. [Accessed 28 November 2016]. |
| [4] | "Small Area Income and Poverty Estimates," United States Census Bureau, 2014. [Online]. Available: https://www.census.gov/did/www/saipe/data/statecounty/data/2014.html. [Accessed 28 November 2016]. |
| [5] | "stackoverflow.com," stackoverflow, 13 December 2016. [Online]. Available: http://stackoverflow.com/. [Accessed 13 December 2016]. |