Cancer Mortality Rates

An Exploratory Data Analysis

Shane Andrade
Suhas Gupta
Ankitkumar Patel

9/22/2018

Introduction

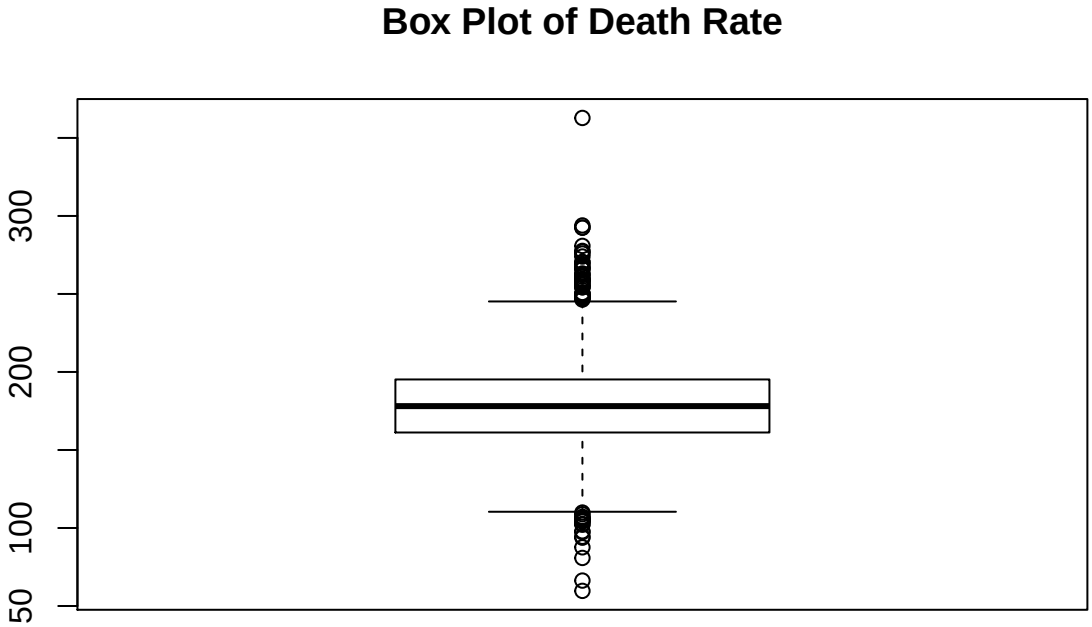
Our motivation for analyzing the provided data set was to provide an understanding of the factors that could contribute to cancer mortality rates. Using this knowledge, the goal would be to provide targeted actions against these factors.

Data Overview

We begin by loading the data from the csv into a data frame and inspecting it. We have a total of 3047 observations and 30 variables:

`X`, `avgAnnCount`, `medIncome`, `popEst2015`, `povertyPercent`, `binnedInc`, `MedianAge`, `MedianAgeMale`, `MedianAgeFemale`, `Geography`, `AvgHouseholdSize`, `PercentMarried`, `PctNoHS18_24`, `PctHS18_24`, `PctSomeCol18_24`, `PctBachDeg18_24`, `PctHS25_Over`, `PctBachDeg25_Over`, `PctEmployed16_Over`, `PctUnemployed16_Over`, `PctPrivateCoverage`, `PctEmpPrivCoverage`, `PctPublicCoverage`, `PctWhite`, `PctBlack`, `PctAsian`, `PctOtherRace`, `PctMarriedHouseholds`, `BirthRate`, `deathRate`

We see mostly numerical types, except `Geography` which is categorical and `binnedInc` which is also categorical but represents a histogram bin range for income. `X` appears to be a row identifier and will not be analyzed. `deathRate` is the target variable.



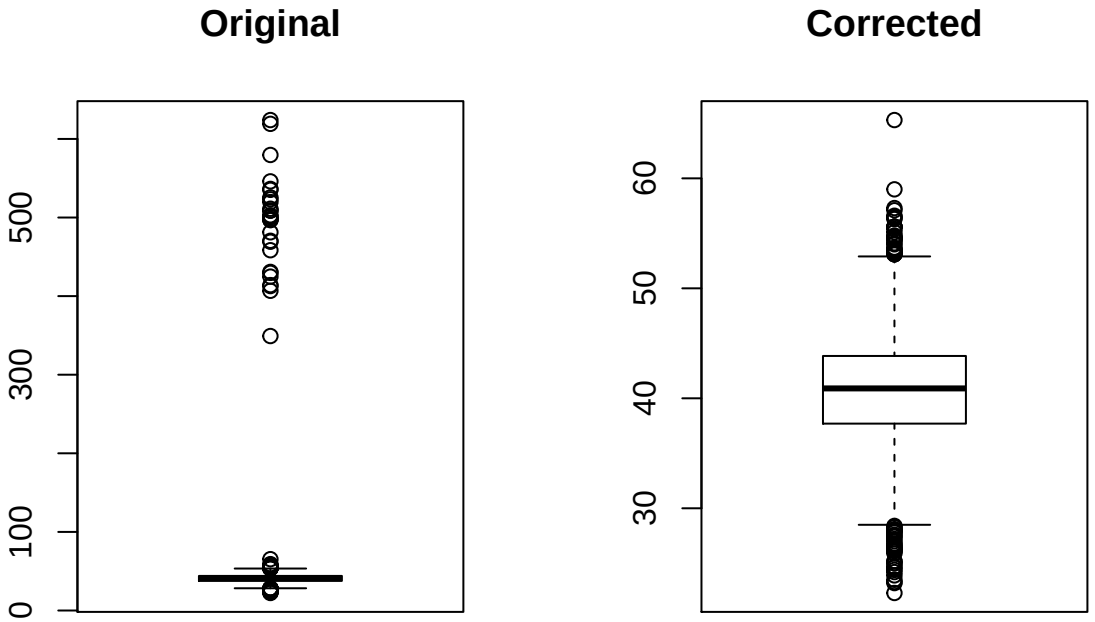
Looking at the mean and median of the `deathRate` variable, 178.66 and 178.10 respectively, we see there is no skew. The histogram above confirms this with a fairly even distribution of the data. The box plot shows this as well, however we do see the presence of outliers on the positive and negative sides.

Quality and Preprocessing

The `AvgHouseholdSize` variable has a 61 values less than 1. Since this is not possible, we will set those values to `NA`.

Looking at the `MedianAge` variable, we can see that there are age values that are not possible.

There were a total of 30 records from the data set that have a `MedianAge` value of more than 100. These values have been replaced with the average of `MedianAgeFemale` and `MedianAgeMale`. The value 100 was chosen because it is still possible and the next highest value was 349.2. Below are the box plots of the original and corrected `MedianAge` values.



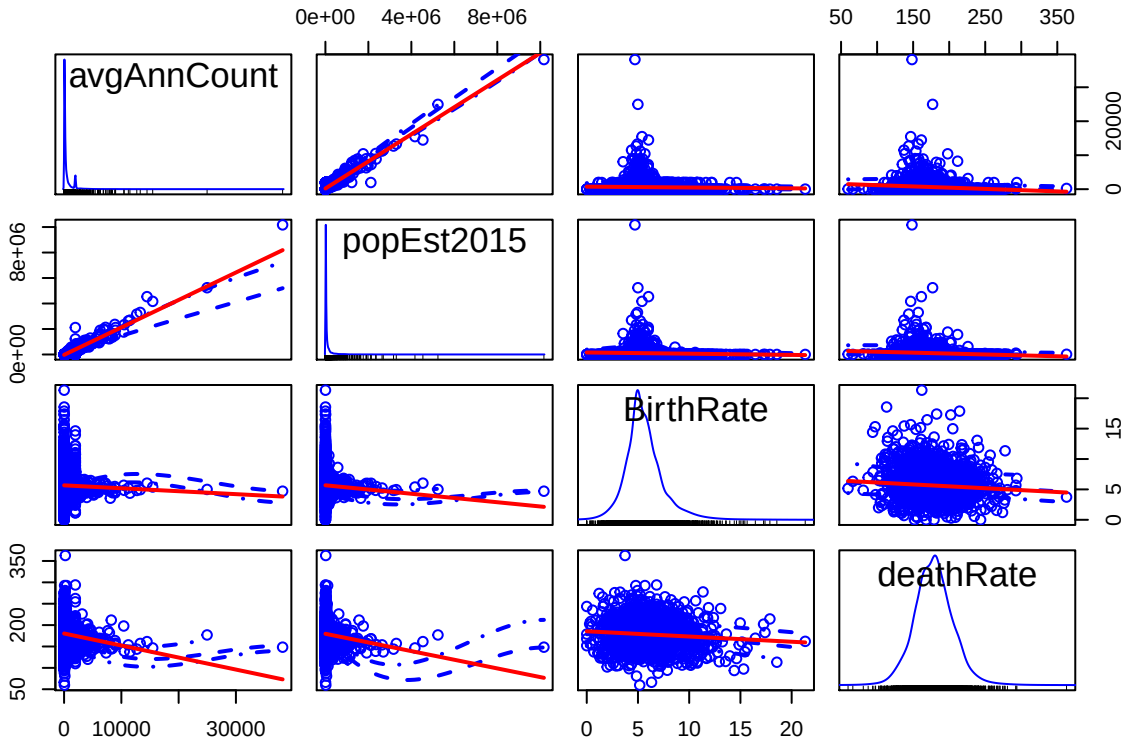
The `PctSomeCol18_24` field has 74.99% data that is missing, which can influence the data analysis. Thus, we excluded this variable from this analysis. On the other hand, only 4.99% of data missing for `PctEmployed16_Over`, which may not have much influence on the statical properties, and thus, we consided the parameter and its effects in this analysis.

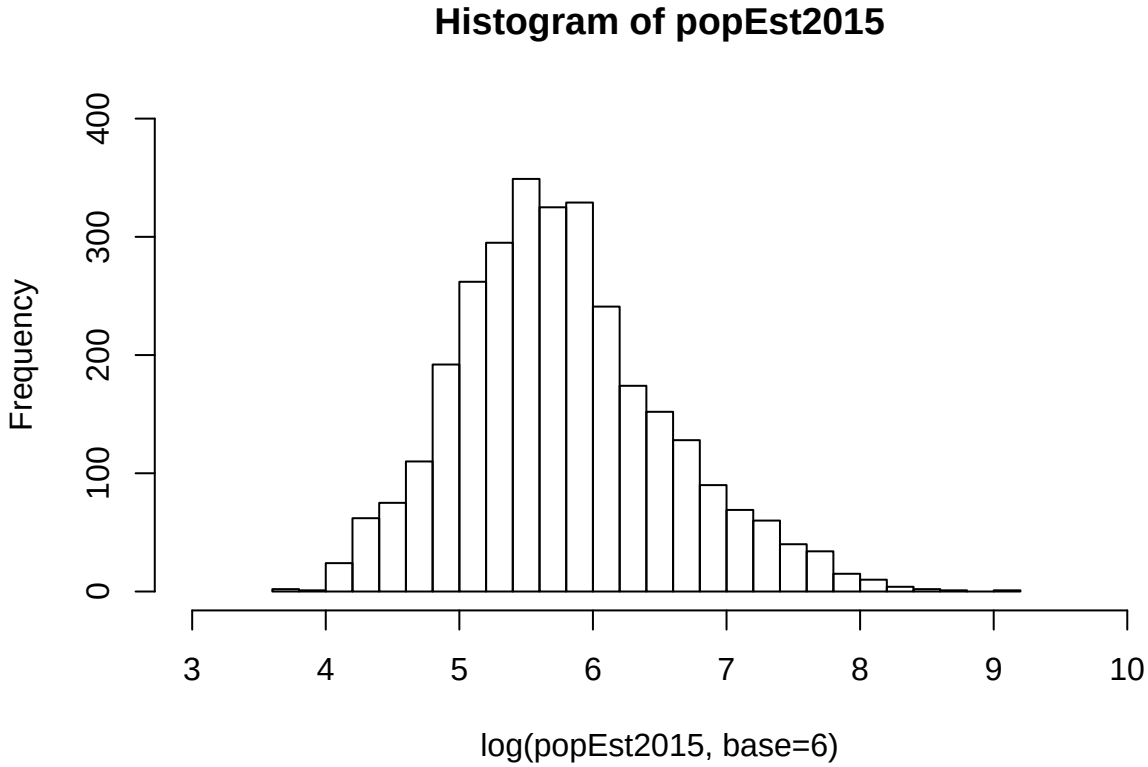
Data Analysis

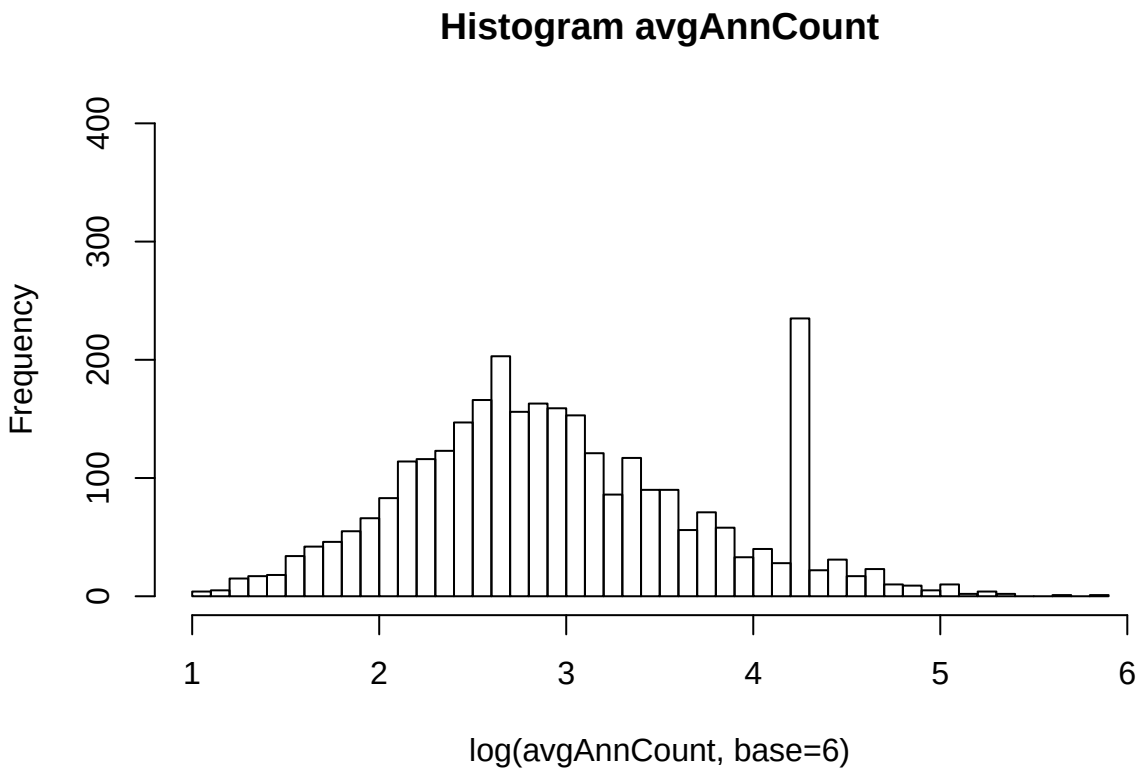
Since there are a large number of variables in the data set, we will use a scatterplot matrix to get a high level overview of the key relationships. As there are several related groups of socioeconomic variables, we will categorize related variables and analyze them together as a group.

County Population

The variables associated with the county population are `avgAnnCount`, `popEst2015`, and `BirthRate`.

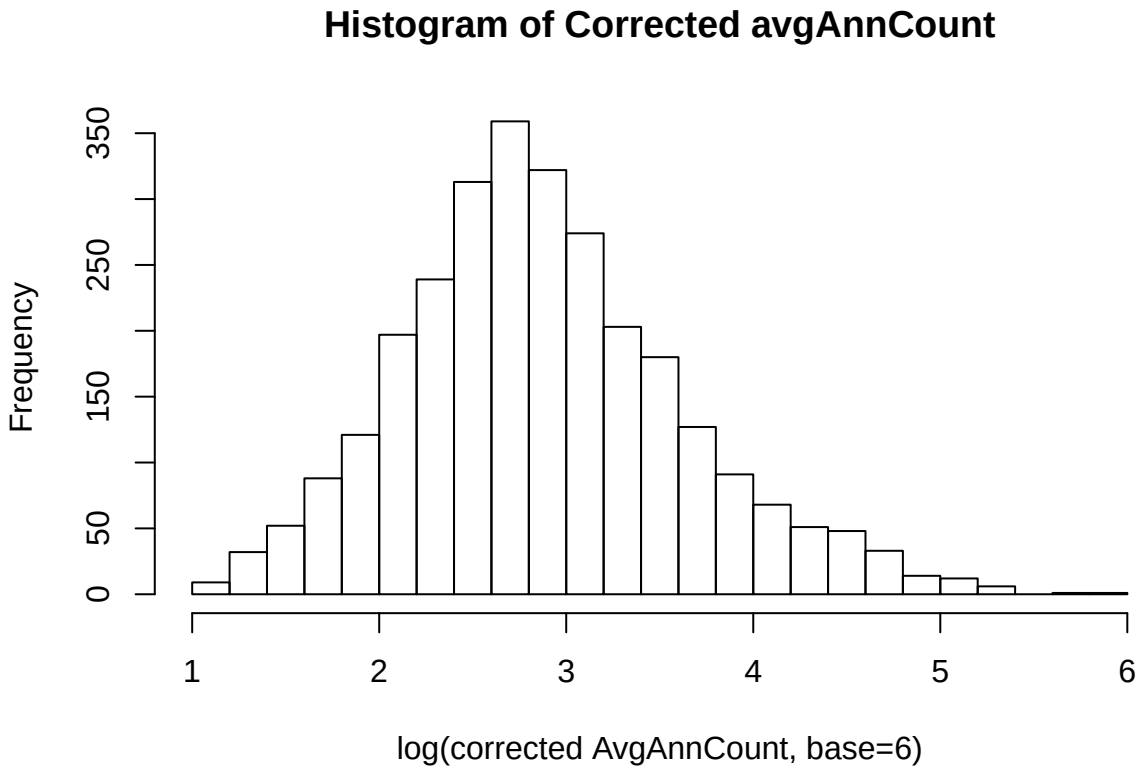

avgAnnCount
popEst2015
BirthRate
deathRate
0e+00
4e+06
8e+06
50
150
250
350
0
20000
0e+00
8e+06
0
5
15
50
200
350
0
10000
30000
0
5
10
15
20


Histogram of popEst2015
Frequency
0
100
200
300
400
3
4
5
6
7
8
9
10
log(popEst2015, base=6)



We observe an extreme positive skew with the variables `popEst2015` and `avgAnnCount` in the histograms above. We apply the log transformation to the variables `avgAnnCount` and `popEst2015`. We use 6 as the base of the log since the minimum value of `avgAnnCount` is 6.00.

Exploring the data in `avgAnnCount` more, we find that values that show the second maximum in `avgAnnCount` appear to be copies of the same values `1962.678`. We think it is highly unlikely to have average incidences of cancer over 4 years to be exactly the same value for multiple counties. Hence, we conclude that these must be erroneous values. We suspect that there might be errors during survey data collection and/or recording. We will replace the `avgAnnCount` column for these values with `NA`. There are a total of 206 affected records. Below is the histogram after the erroneous values are replaced with `NA`.

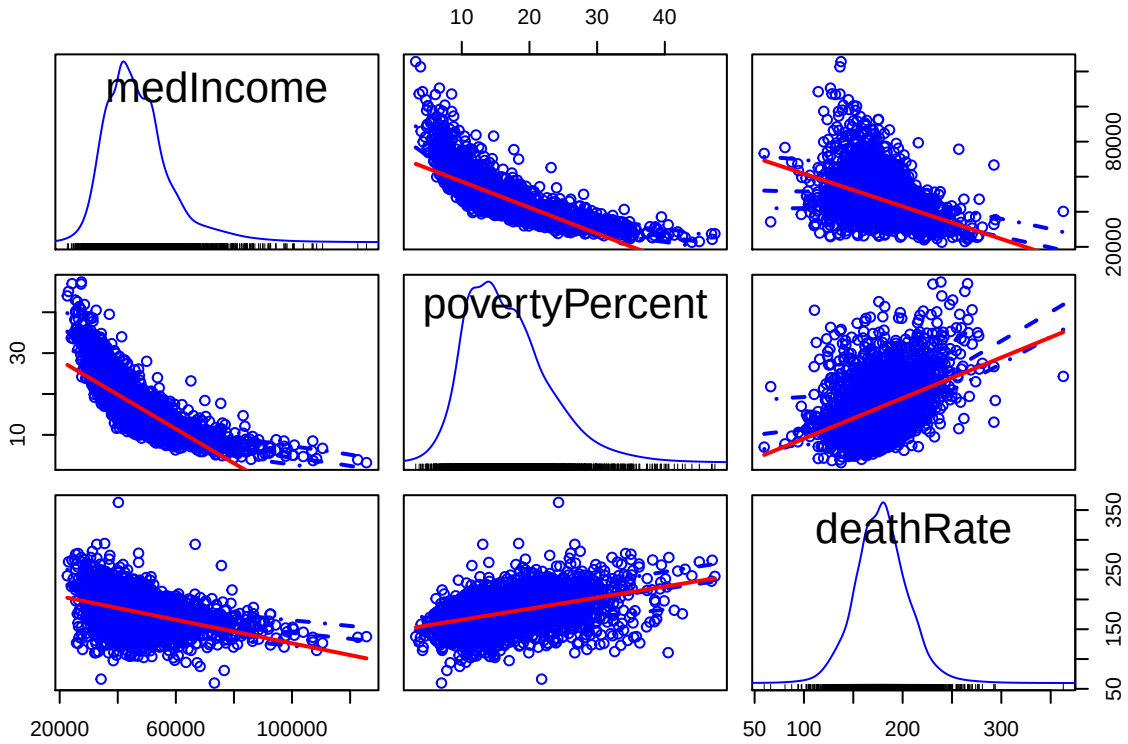


avgAnnCount	popEst2015	BirthRate
-0.1435	-0.1201	-0.0874

The correlations of the corrected population variables against `deathRate` are shown above. The population variables have a small negative correlation with the cancer death rate.

Income

The variables associated with income are `medIncome`, `povertyPercent`, and `binnedInc`.

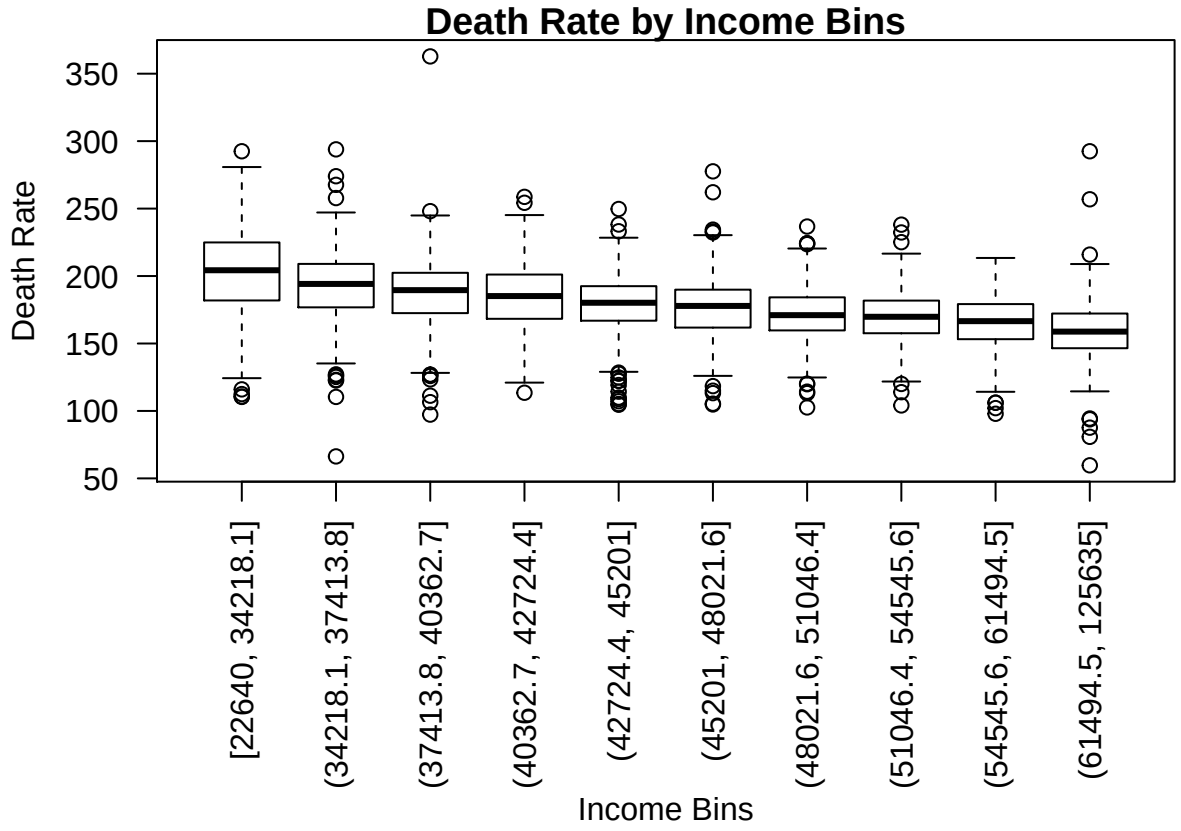


The scatterplot matrix shows a strong linear relationships with income variables. The table below confirms the correlation of `deathRate` with `medIncome` and `povertyPercent`.

medIncome	povertyPercent
-0.4286	0.4294

As one might expect, the amount of poverty and the median income are correlated to the cancer death rate. We can see, as poverty increases, the death rate also increases and when the median income increases the death rate decreases.

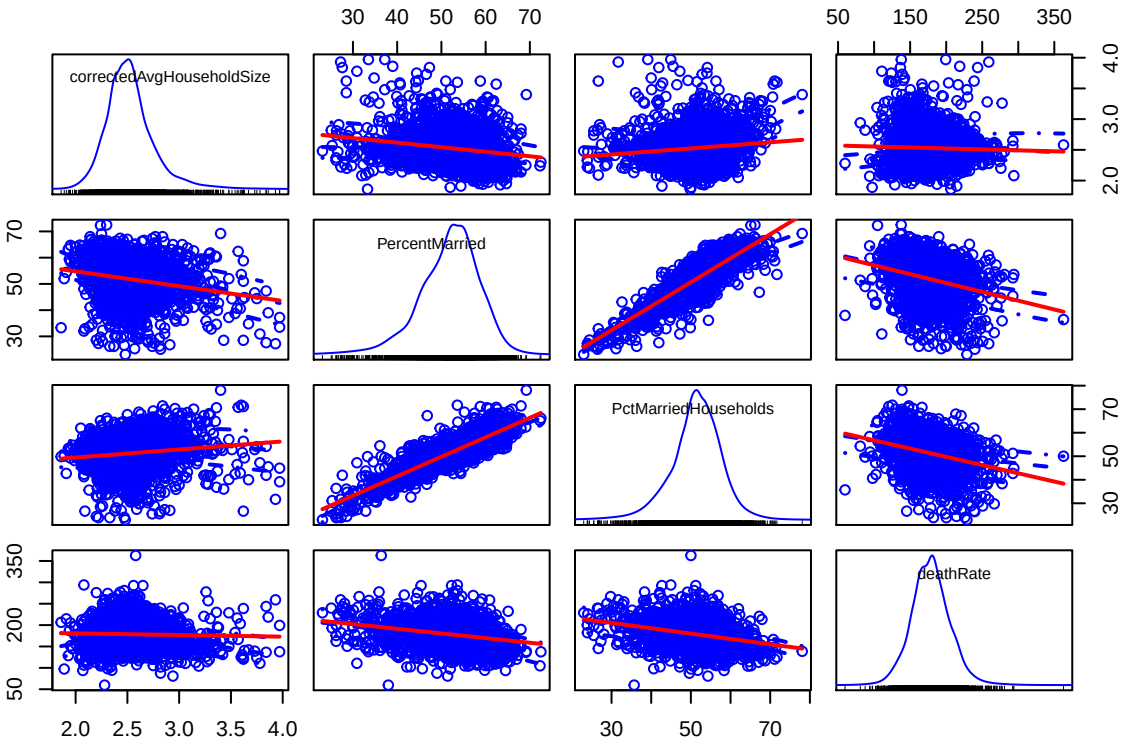
Since the `binnedInc` was a categorical variable, we decided to analyze it separately. Below is a categorical boxplot of each of the `binnedInc` values against death rate.



The death rate has a negative linear trend with increasing income bins.

Household

The following variables have been determined to be considered as describing a household: `AvgHouseholdSize`, `PercentMarried`, `PctMarriedHouseholds`, so they will be analyzed together along with the target variable `deathRate`. Note that this is excluding the `NA` values we introduced to `AvgHouseholdSize`.



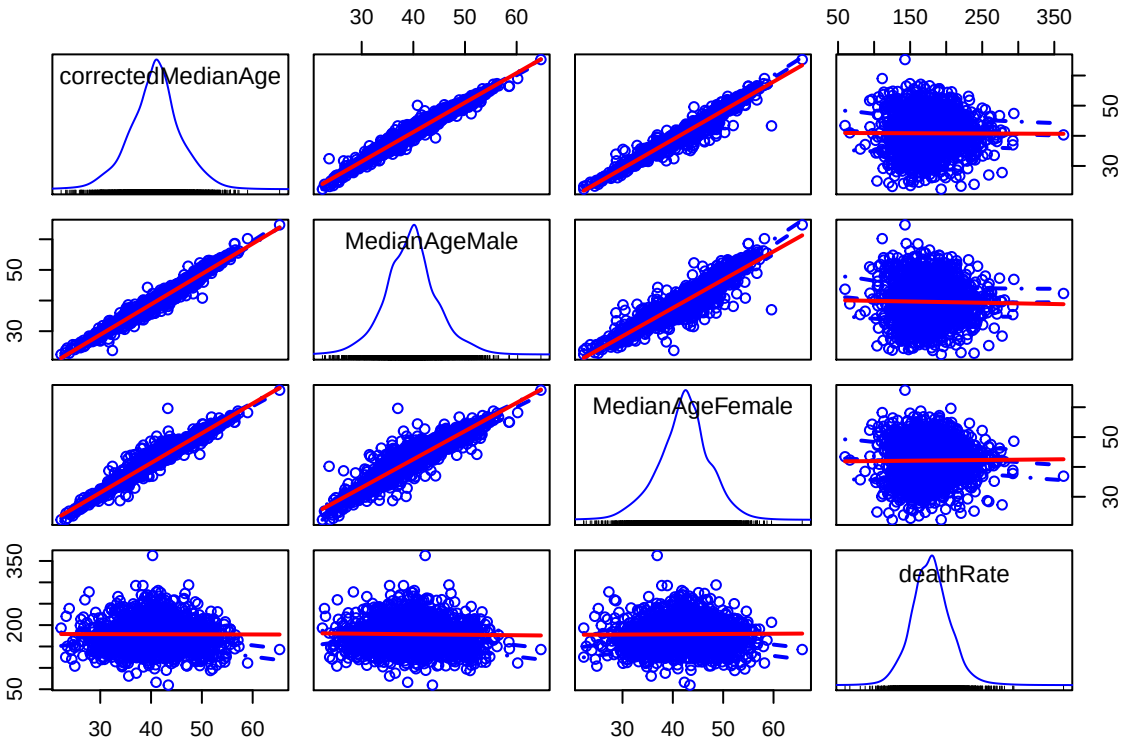
In the bottom row of the scatterplot matrix above, we can see the correlations of each variable against the `deathRate`. They are summarized numerically in the table below:

`correctedAvgHouseholdSize`	`PercentMarried`	`PctMarriedHouseholds`
-0.0346	-0.2668	-0.2933

Marriage appears to have a negative correlation to cancer death rate. Comparatively, the size of the household itself does not have a significant correlation.

Age

The variables related to age are `MedianAge`, `MedianAgeMale`, and `MedianAgeFemale`. The following is a scatterplot matrix of these variables with the `deathRate` variable. Note that this is using the corrected data for `MedianAge`.



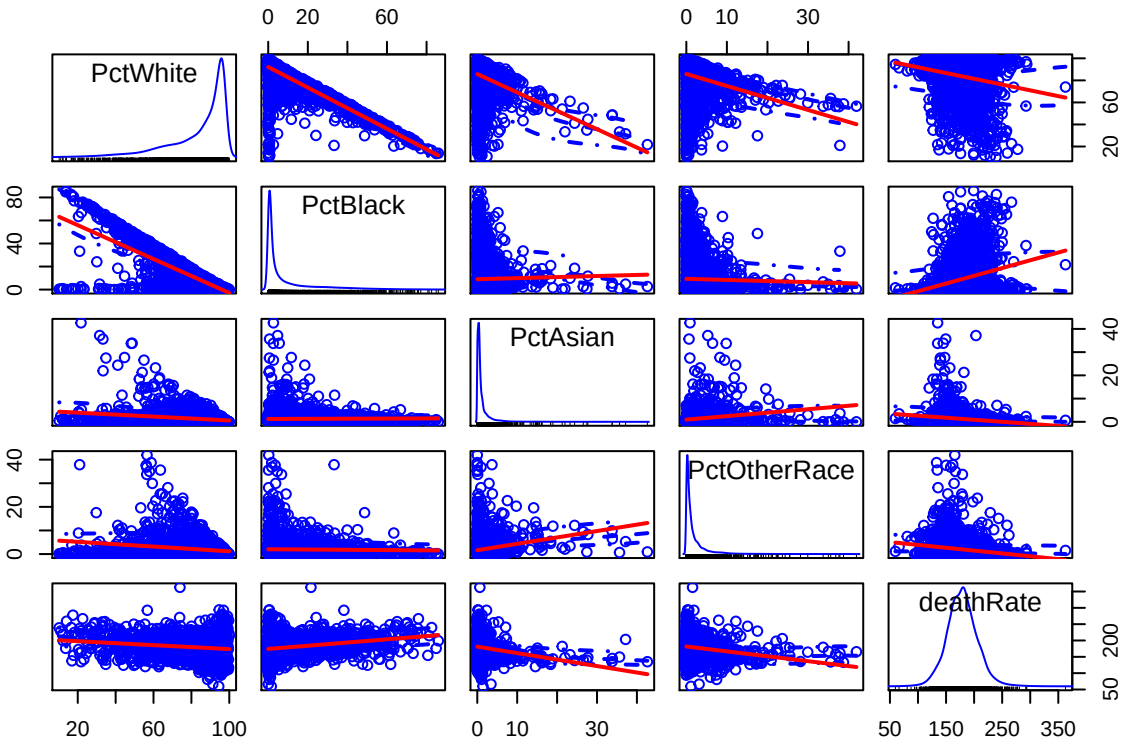
The correlations of the variables to the `deathRate` can be found in the bottom row but appear to have little or no correlation.

`MedianAge`	`MedianAgeFemale`	`MedianAgeMale`
-0.0049	0.0120	-0.0219

One interesting note here is that we see a positive correlation in the median female age to cancer death rates but a negative one in median male ages. Our data set does not contain any other gender demographics, but if it did it would be interesting to see if that has any correlation to cancer death rate.

Race

The variables `PctWhite`, `PctBlack`, `PctAsian`, and `PctOtherRace` have been identified as dealing with county demographic data pertaining to race. Below is a scatterplot matrix of the race variables and their correlation to `deathRate`.

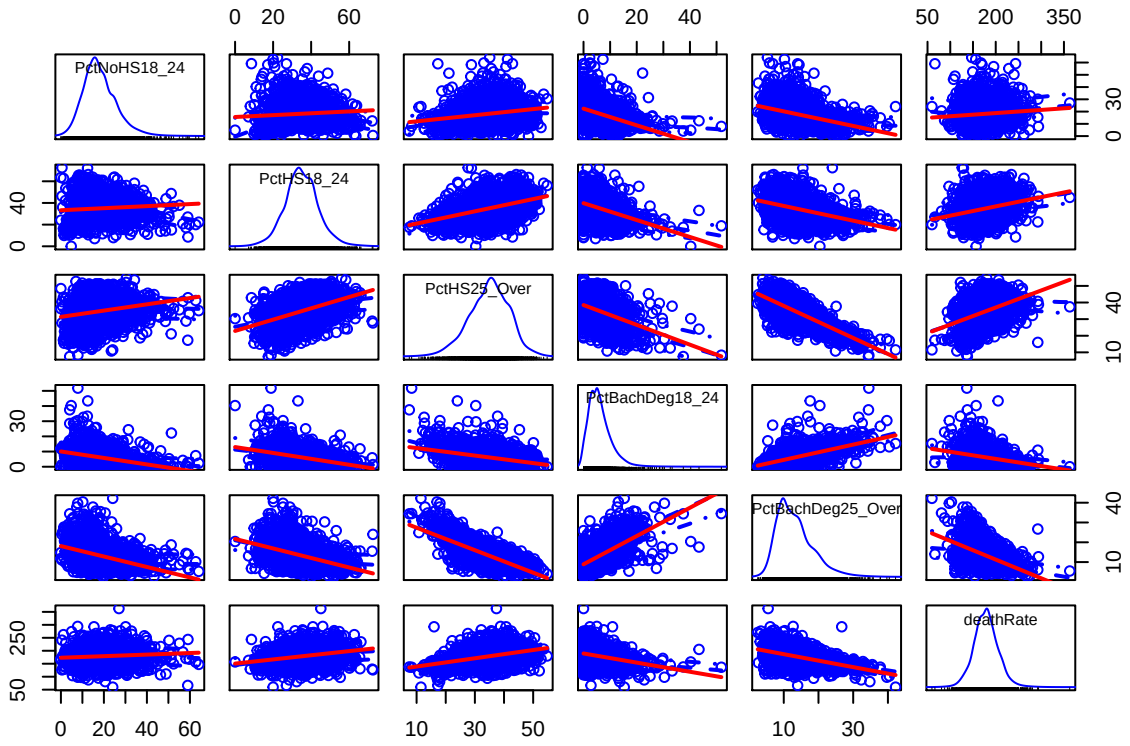


PctWhite	PctBlack	PctAsian	PctOtherRace
-0.1774	0.2570	-0.1863	-0.1899

Racial demographics are highly correlated with cancer death rate. The death rate increases with the percentage of black population in a county; however, the death rate decreases as white, asian, and other population increases.

Education

The variables that capture education in a county are `PctNoHS18_24`, `PctHS18_24`, `PctBachDeg18_24`, `PctHS25_Over`, and `PctBachDeg25_Over`.



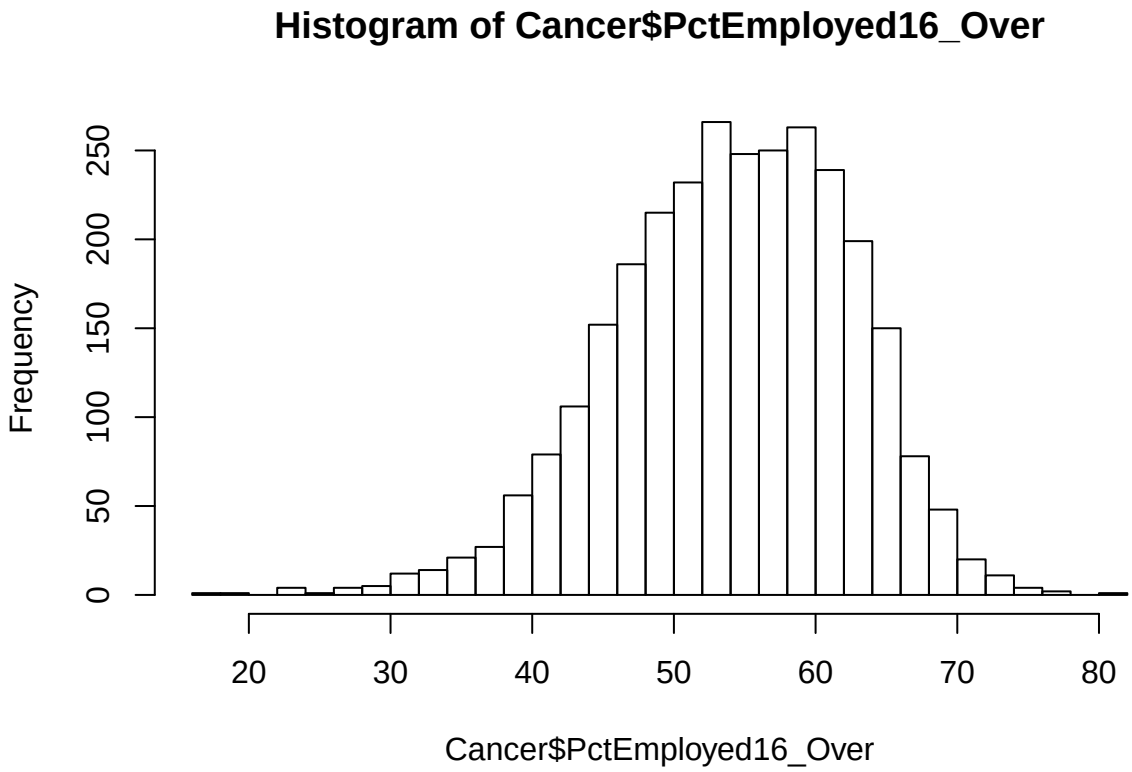
PctNoHS18_24	PctHS18_24	PctHS25_Over	PctBachDeg18_24	PctBachDeg25_Over
0.0885	0.2620	0.4046	-0.2878	-0.4855

The level of education appears to relate to the cancer death rate. Those with no college education have a higher correlation of cancer mortality. Pursuing a some college degree seems to have a negative correlation on cancer mortality. Futhermore, the age at which you receive a bachelors degree is even more negatively correlated to the cancer mortality rate. This is the opposite to high school education which is more positively correlated to cancer death rates the later in life the high school diploma is received.

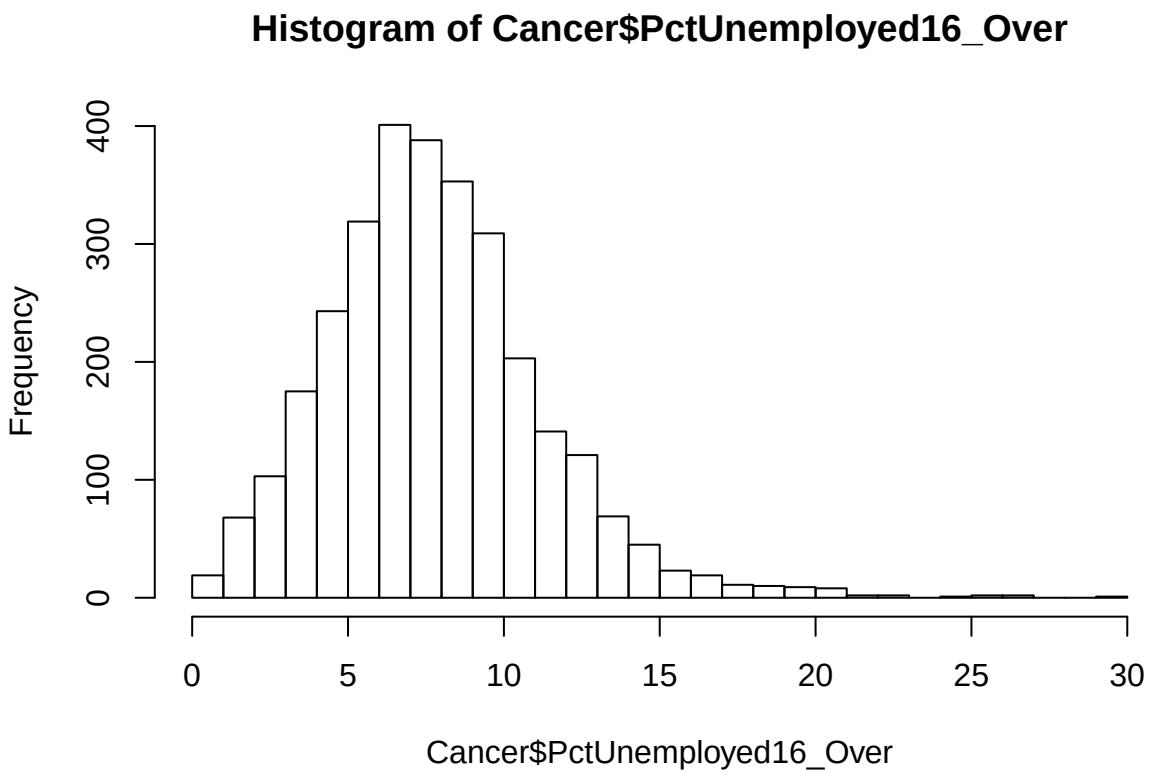
Employment

There are two employment variables in the data set whose histograms are shown below:

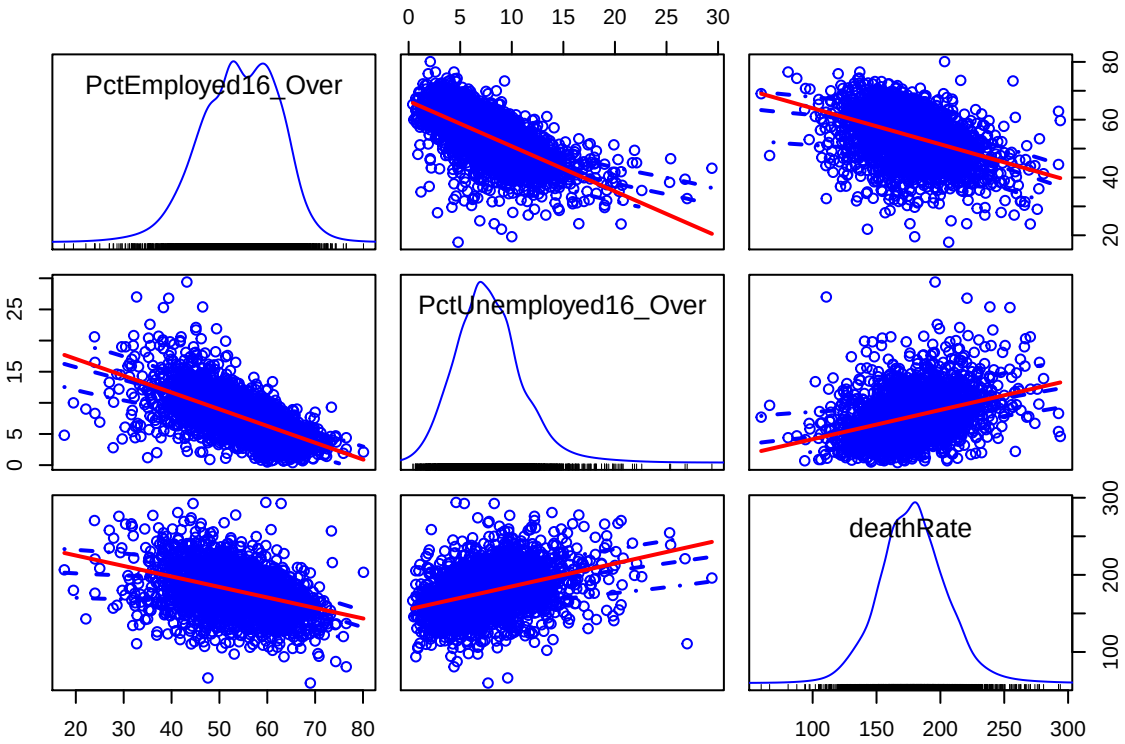
```
hist(Cancer$PctEmployed16_Over, breaks=25)
```




```
hist(Cancer$PctUnemployed16_Over, breaks=25)
```



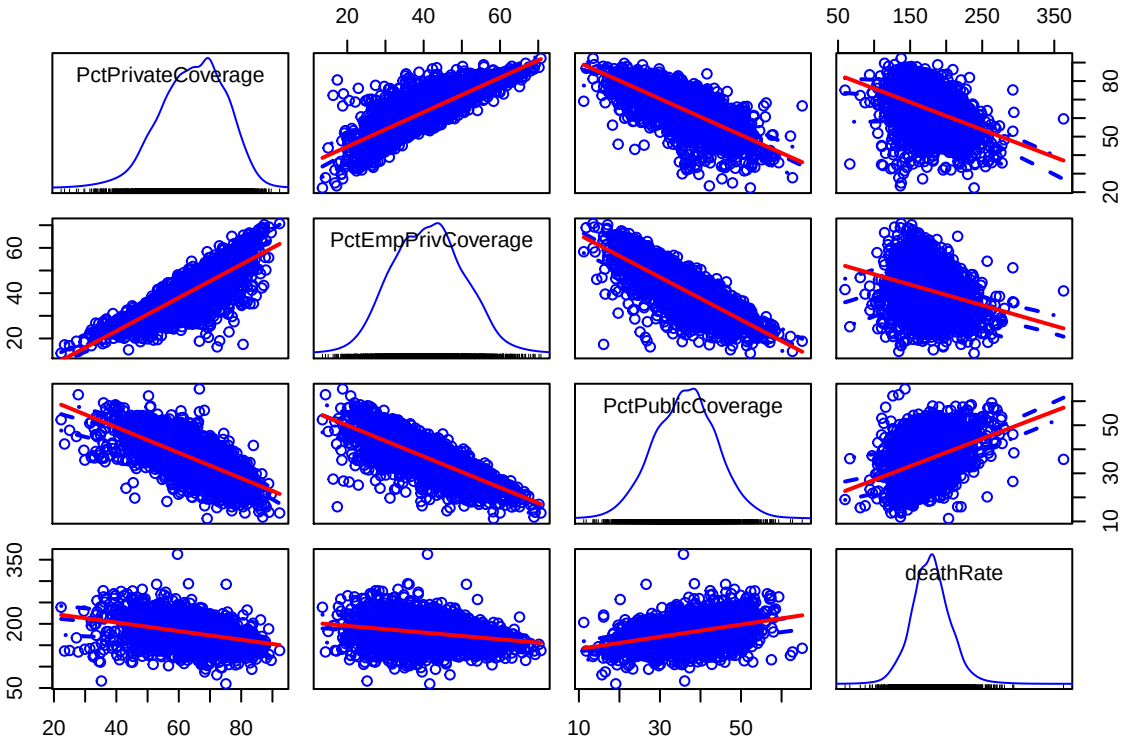
We observe in the histograms that the percentages of employed and unemployed have a negative and positive skew repectively.



|—:|—:| |NA|0.3784|

Healthcare Coverage

The following variables are related to healthcare coverage: `PctPrivateCoverage`, `PctEmpPrivCoverage`, and `PctPublicCoverage`.



PctPrivateCoverage	PctEmpPrivCoverage	PctPublicCoverage
-0.3861	-0.2674	0.4046

The death rate increases with the percentage of population with public insurance coverage. On the other hand, the death rate decreases with as the percentage of population with private or employee-sponsored insurance coverage increases in a county.

Additional data that could improve analysis

1. We found that for certain counties the "avgAnnCount" variables values were larger than the "popEst2015" values. we have included data corresponding to these counties in our analysis since we do not have information on the relationship between these variables. It is possible that the actual relationship bewtween these variables requires additional filtering of the data set and the current correlation analysis might be impacted by this missing information.
2. Types of food consumption: —
3. Family history of disease: —

Confounding Variables

There are some variables for which data are not included in the data set used for our analysis. However, these variables (summarized below) could have an impact on the distribution of the independant variables that we see in the data set. This could in turn lead to the relationship that we observe for death rate. Data on these variables would help in understading the trend of the dependant variable better.

1. Crime Rate
2. Weather
3. Housing Prices

The above variables could affect what races, income and household types live in certain counties. Thus these could be indirectly affecting the variable death rate through the independant variables that we have analyzed in our study.

Conclusion

With the goal of identifying communities to target for intervention through the above statistical analysis, we have summarized the key inferences below:

1. The strongest factors increasing the death rate are poverty percentage, public heath coverage, high school only education, unemployment and percentage of black population (in decreasing order of correlation)
2. The strongest factors decreasing the death reate are increasing median income, private health coverage, bachelors degree and marital status (in decreasing order of correlation)

The above information can be used to target counties effectively with an aim to reduce the number of deaths from cancer.