Lab 1: Cancer EDA

Daniel Rasband, Subha Vadakkumkoor, Hong Yang

January 30, 2018

Introduction

Research Question

We were hired by a health government agency to understand and predict cancer mortality rates:

1. Understand factors that predict cancer mortality rates
2. Identity communities for social interventions
3. Understand which interventions are likely to have impact

Objective of this Lab

Our task at hand is to perform an exploratory data analysis to understand how county-level characteristics are related to cancer mortality.

Data Set

Read the provides CSV file, cancer.csv file into an R dataframe `canc.dat`.

```
canc.dat <- read.csv('cancer.csv', header = TRUE)
```

Number of Observations, variables, datatypes and granularity

There are 3047 observations and 30 columns in the data set. All columns are numeric or integer except Geography and binnedInc that are factors.

```
dim(canc.dat)
```

```
## [1] 3047   30
```

```
sapply(canc.dat, class)
```

```
##                     X          avgAnnCount            medIncome
##             "integer"            "numeric"            "integer"
##           popEst2015       povertyPercent            binnedInc
##             "integer"            "numeric"             "factor"
##            MedianAge        MedianAgeMale      MedianAgeFemale
##            "numeric"            "numeric"            "numeric"
##            Geography     AvgHouseholdSize       PercentMarried
##             "factor"            "numeric"            "numeric"
##         PctNoHS18_24           PctHS18_24      PctSomeCol18_24
##            "numeric"            "numeric"            "numeric"
##      PctBachDeg18_24         PctHS25_Over    PctBachDeg25_Over
##            "numeric"            "numeric"            "numeric"
##   PctEmployed16_Over PctUnemployed16_Over   PctPrivateCoverage
##            "numeric"            "numeric"            "numeric"
##   PctEmpPrivCoverage    PctPublicCoverage             PctWhite
##            "numeric"            "numeric"            "numeric"
```

```
##             PctBlack             PctAsian         PctOtherRace
##            "numeric"            "numeric"            "numeric"
## PctMarriedHouseholds            BirthRate            deathRate
##            "numeric"            "numeric"            "numeric"
```

The number of counties equals the number of observations, so no county has multiple rows of data. This confirms that the granularity of the data is at county level.

```
length(unique(canc.dat$Geography)) == nrow(canc.dat)
```

```
## [1] TRUE
```

The data covers 3047 counties from all US states and the District of Columbia. As of 2016, there were 3,144 counties or county equivalents[1], and this data set covers almost all of them.

```
# Split out the county name and state.
geo.dat <- as.character(canc.dat$Geography)
split.geodata <- str_split_fixed(canc.dat$Geography, ', ', 2)
canc.dat$county <- split.geodata[ , 1]
canc.dat$state <- split.geodata[ , 2]
length(unique(canc.dat$state))
```

```
## [1] 51
```

Data Quality and Preparation

Data Overview

Let us summarize the fields to look for distribution, missing values, range and outliers and to understand the fields better. The X column appears to be a simple auto-incremented ID, so it is ignored. Rest of the county variables can be broadly classified as belonging to the following groups:

- Census:
  - `popEst2015`: Estimated population in 2015
  - `BirthRate`: The birth rate, unknown basis.
  - `deathRate`: The death rate, per 100,000 people.
  - `avgAnnCount`: Average cancer total incidences per year from 2009-2013.
  - `Geography`: County, State
- Age:
  - `MedianAge`: Median Age
  - `MedianAgeMale`: Median Age of Males
  - `MedianAgeFemale`: Median Age of Females
- Income:
  - `medIncome`: Median Income
  - `povertyPercent`: Percentage of population below poverty level
  - `binnedInc`: Income range
- Household:
  - `AvgHouseholdSize`: Average Household Size
  - `PercentMarried`: Percent of population that is married
  - `PctMarriedHouseholds`: Percentage of the population that is in a married household.
- Race:
  - `PctWhite`: Percentage of population that is white.
  - `PctBlack`: Percentage of population that is black.
  - `PctAsian`: Percentage of population that is asian.
  - `PctOtherRace`: Percentage of population that is not white, black, or asian.

[1] "County (United States)." Wikipedia, Wikimedia Foundation, 24 Jan. 2018, en.wikipedia.org/wiki/County_(United_States).

- Education:
  - `PctNoHS18_24`: Percent of population, ages 18-24, that hasn't graduated from high school.
  - `PctHS18_24`: Percent of population, ages 18-24, that has graduated from high school.
  - `PctSomeCol18_24`: Percentage of 18-24-year-olds that have some college education.
  - `PctBachDeg18_24`: Percentage of 18-24-year-olds that have a bachelor's degree.
  - `PctHS25_Over`: Percentage of population, 25 and older, graduated from high school.
  - `PctBachDeg25_Over`: Percentage of population, 25 and older, graduated from college.
- Employment:
  - `PctEmployed16_Over`: Percentage of population, 16 and older, that is employed.
  - `PctUnemployed16_Over`: Percentage of population, 16 and older, that is unemployed.
- Insurance:
  - `PctPrivateCoverage`: Percentage of population with private insurance coverage.
  - `PctEmpPrivCoverage`: Percentage of population with private insurance coverage from employment.
  - `PctPublicCoverage`: Percentage of population with public insurance coverage.

The most questionable of the above assumptions is that of the `deathRate`, but comparing the numbers with those in a PDF of death-related statistics for Los Angeles county[2], and by using a bit of common sense, it appears that the rate is per 100,000 people. The birth rate is similarly questionable, but by using Los Angeles's data, it appears that the number may be per 100 women, ages 15 to 44, though this is not certain.

Another fundamentally questionable aspect of the data is that most data points do not have a reference year. We are assuming 2015 for those data, but that may not be accurate. For example, `avgAnnCount` has been defined as 2009-2013 mean incidences per county while `popEst2015` is defined as of 2015. We will assume that cancer incidents in 2015 are similar to that of 2009-2013, which is a big assumption that can impact the robustness of the analysis.

Although the analysis relates to cancer, there is no data about the availability of hospitals or oncology specialization that are related to if and how the cancer is treated. There are no also details on the nature of cancer patients like age or gender or type of the disease, or other known cancer causing habits like smoking.

As lack of data that directly or indirectly pertains to cancer will weaken the analysis and impact the outcome, a request to make these data elements available should be placed with the government agency.

Looking at the range of `popEst2015`, `medIncome` and `povertyPercent`, we can see that a range of counties from small to big, rich and poor have been included in the dataset. `AvgHouseholdSize` varies from 0.02 to 3.9 which indicates both big and small families have been included.

While the distribution of `PctWhite` looks even, `PctBlack` and `PctAsian` have outliers. While ~75% of the counties have a black population of 10.5% , some others have as high as 85%. Similarly, 75% of communities have about 1% of Asian population, but some others have 43%. We have different types of neighborhoods in the data, and some neighborhoods that are very different from others.

Univariate Analysis

Data Issues

There are various issues with the data that require us some massaging for further analysis.

The `PctSomeCol18_24` variable has a large majority of missing values, so we are opting to remove the column entirely for the dataset as the data is too sparse to use.

```
# Remove PctSomeCol18_24 completely.
fixed.dat <- subset(canc.dat, select = -c(PctSomeCol18_24))
'PctSomeCol18_24' %in% names(fixed.dat)
```

[2]The Office of Health Assessment and Epidemiology. "Mortality in Los Angeles County." 7 Oct. 2015, doi:http://publichealth.lacounty.gov/dca/data/documents/mortalitypresentation2012.pdf.

```
## [1] FALSE
```

Field `PctEmployed16_Over` has 152 rows with missing data, but we will keep this field in the analysis.

The `MedianAgeMale` and `MedianAgeFemale` fields look good but `MedianAge` has a high value of 624.00. It appears that `MedianAge` is set to months, rather than years, in a number of cases, so we adjust those rows:

```
age.outliers <- fixed.dat$MedianAge > 100
fixed.dat$MedianAge[age.outliers] <- fixed.dat$MedianAge[age.outliers] / 12
# Sanity check
summary(fixed.dat$MedianAge, fixed.dat$MedianAgeFemale, fixed.dat$MedianAgeMale)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.30   37.70   40.90   40.83   43.80   65.30
```

The `avgAnnCount` variable has 206 rows that are 1) the exact same number, and 2) greater than the actual population. This indicates to us that those values were missing from the data and had been assigned a default value (`1962.668`). Because the `avgAnnCount` variable is central to this analysis, and those particular values have proven nonsensical, we will have to remove those entire rows:

```
canc.dat[canc.dat$avgAnnCount > canc.dat$popEst2015, c('X', 'avgAnnCount')]
```

```
##         X avgAnnCount
## 116   116    1962.668
## 1299 1299    1962.668
## 1305 1305    1962.668
## 1313 1313    1962.668
## 1362 1362    1962.668
## 3034 3034    1962.668
```

This reduces the number of observations from 3047 to 2841.

```
fixed.dat <- subset(fixed.dat, avgAnnCount != canc.dat$avgAnnCount[116])
nrow(fixed.dat)
```

```
## [1] 2841
```

Let us check if all the percentage fields related to a particular element add upto 100 to see if there are parts of the population not represented in this data and also to help better estimate what the base of the percentage calculation is.

- PctHS18_24 and PctNoHS18_24 do not add upto 100 in almost all cases, just as PctEmployed16_Over and PctUnemployed16_Over do not. This means these are not percentages calculated based on the entire population.

```
nrow(subset(canc.dat, (fixed.dat$PctHS18_24+ fixed.dat$PctNoHS18_24)<100))
```

```
## [1] 3046
```

```
summary(fixed.dat$PctHS18_24+fixed.dat$PctNoHS18_24)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.50   45.80   54.40   53.54   61.80  100.00
```

```
nrow(subset(fixed.dat, (fixed.dat$PctEmployed16_Over + fixed.dat$PctUnemployed16_Over) < 100))
```

```
## [1] 2700
```

```
summary(fixed.dat$PctEmployed16_Over + fixed.dat$PctUnemployed16_Over)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   22.40   57.70   62.10   61.71   66.30   82.70     141
```

- Race fields all add upto 100 indicating these are mutually exhaustive classifications of the data.

```
nrow(subset(canc.dat, (fixed.dat$PctWhite + fixed.dat$PctBlack + fixed.dat$PctAsian + fixed.dat$OtherRa
```

```
## [1] 0
```

- There is a part of the data that have neither public nor private coverage, indicating a set of population who have no insurance. Also the total percents are very often greater than 100 indicating a part of population that have both private and public insurance.

```
nrow(subset(fixed.dat, (fixed.dat$PctPrivateCoverage + fixed.dat$PctPublicCoverage < 100)))
```

```
## [1] 1284
```

```
summary(fixed.dat$PctPrivateCoverage + fixed.dat$PctPublicCoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    65.4    96.0   100.8   100.2   105.3   131.7
```

Target Variable Analysis

Our task is to look for correlations between cancer mortality rate and other factors, but cancer mortality rate is not given. This is the biggest issue with the data and our assumption here greatly reduces the effectiveness of the exploratory data analysis. Our effort here will be to compute the rate of deaths to cancer incidents, and use that as a relative indicator for how frequently cancer causes death amongst the different counties.

Here we calculate the percentage of the population that had cancer. This may be flawed, because some people may have been diagnosed twice or more.

```
fixed.dat$cancerRate = fixed.dat$avgAnnCount / fixed.dat$popEst2015
summary(fixed.dat$cancerRate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.001403 0.004747 0.005532 0.005507 0.006283 0.014048
```

Cancer incidents are very strongly correlated with population, which means that the more people, the more cancer incidents.

```
cor(fixed.dat$avgAnnCount, fixed.dat$popEst2015)
```

```
## [1] 0.9813224
```

As we believe that deathRate is per 100,000 people, here we calculate the death rate for the county's population and we will call this realDeathRate.

```
# The percentage of each population that died.
fixed.dat$realDeathRate<- (fixed.dat$popEst2015 * fixed.dat$deathRate / 100000)/fixed.dat$popEst2015
summary(fixed.dat$realDeathRate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000597 0.001624 0.001793 0.001796 0.001960 0.003628
```

Although we saw a strong correlation between the cancer rate and population, we do not see the same between death rates and population. This indicates that there are several causes of death that vary noticeably by county. This weakens our efforts to get to cancer mortality rate.

```
cor(fixed.dat$deathRate, fixed.dat$popEst2015)
```

```
## [1] -0.1304156
```

In the absence of accurate mortality rates due to cancer, we calculate our best possible approximation as the rate of deaths to cancer. In other words estimate how many death occur per cancer incidence. It ranges from

0.15 to 1.45, which is a very wide range. Also there are counties with our estimated cancer mortality rate greater than 1, indicating more deaths than cancer, which goes back to having the cause of death unavailable.

```
fixed.dat$deathToCancerRate <- fixed.dat$realDeathRate / fixed.dat$cancerRate
summary(fixed.dat$deathToCancerRate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1537  0.2859  0.3330  0.3407  0.3830  1.4490
```

Analysis of Key Relationships

Now that we have looked at each variable, cleaned the data and defined our target variable, let us examine bivariate relationships. We will look at certain groups of independant variables at a time.

Income, Poverty & Insurance Coverage

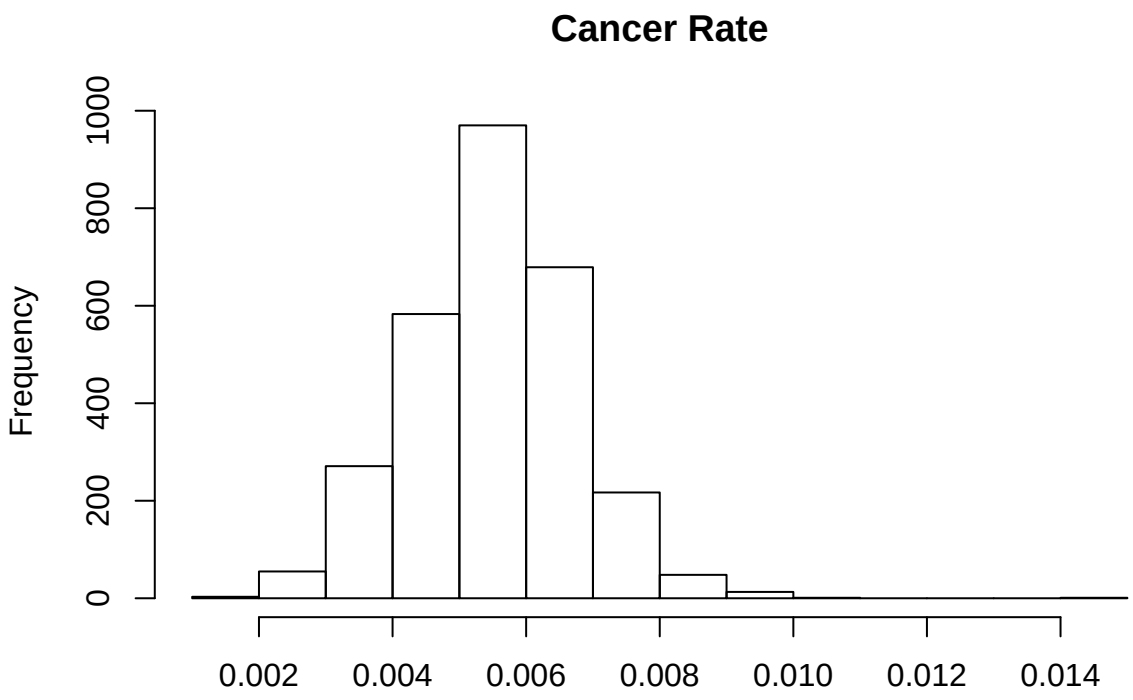
When cancerRate is high, PctPublicCoverage is high and medIncome is relativly low, we will further examine them. We also noticed that public coverage and private coverage seem mutually exclusive.

```
cor(fixed.dat[ , c("cancerRate","PctPrivateCoverage","PctPublicCoverage","povertyPercent", "medIncome",'
```

```
##                     cancerRate PctPrivateCoverage PctPublicCoverage
## cancerRate          1.000000000        0.002481271         0.4927476
## PctPrivateCoverage  0.002481271        1.000000000        -0.7224096
## PctPublicCoverage   0.492747640       -0.722409606         1.0000000
## povertyPercent     -0.003797082       -0.819348715         0.6551984
## medIncome          -0.278654584        0.728851394        -0.7559027
## PctEmpPrivCoverage -0.228859552        0.834285327        -0.7757656
##                    povertyPercent  medIncome PctEmpPrivCoverage
## cancerRate           -0.003797082 -0.2786546         -0.2288596
## PctPrivateCoverage   -0.819348715  0.7288514          0.8342853
## PctPublicCoverage     0.655198356 -0.7559027         -0.7757656
## povertyPercent        1.000000000 -0.7915140         -0.6850845
## medIncome            -0.791513965  1.0000000          0.7434347
## PctEmpPrivCoverage   -0.685084485  0.7434347          1.0000000
```

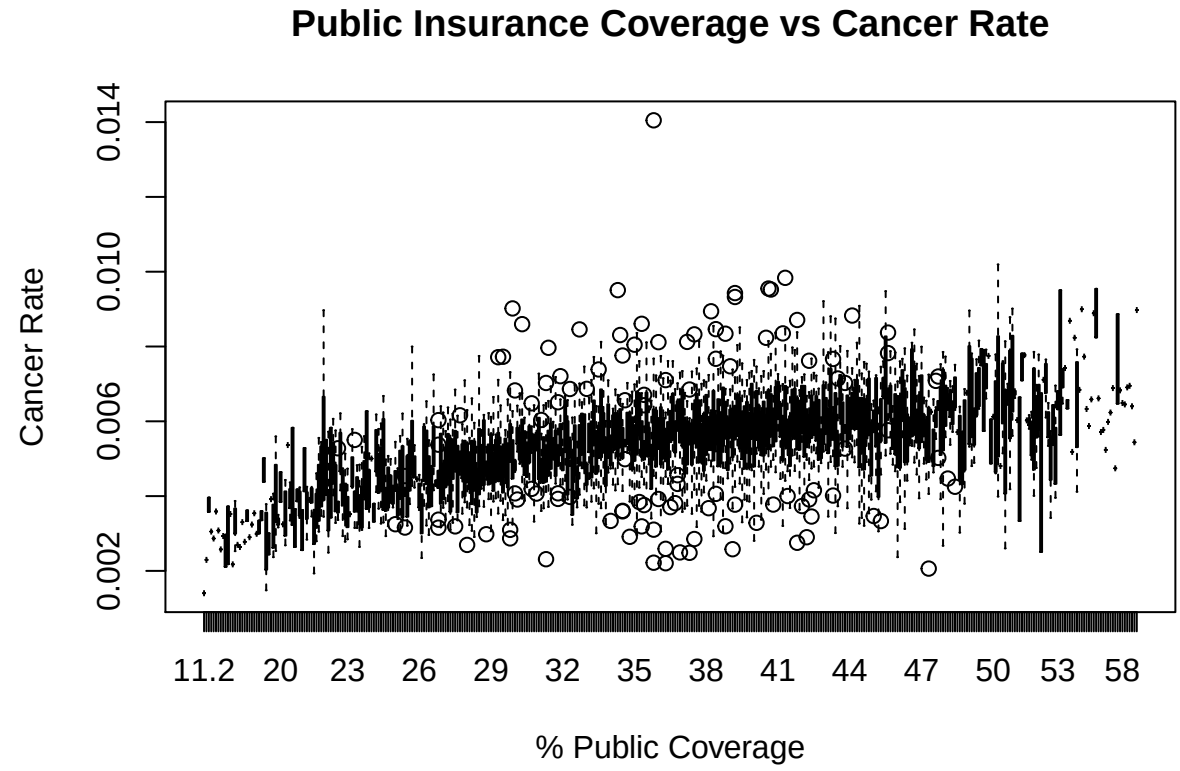
This initial examination shows that `PctPublicCoverage` and `medIncome` are correlated with `cancerRate`. Other variables don't have much strong impact from these plots.

```
hist(fixed.dat$cancerRate, main = "Cancer Rate", xlab = NULL)
```



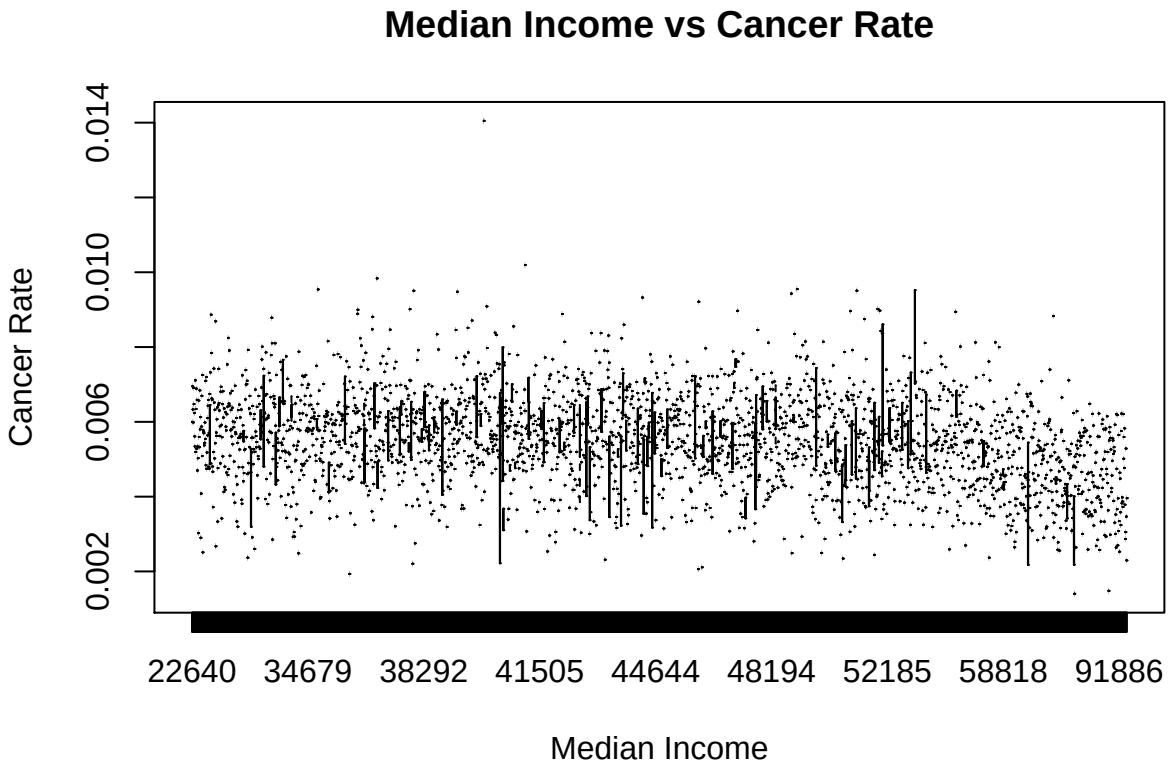
Public insurance coverage is somewhat correlated with the cancer rate.

```
boxplot(fixed.dat$cancerRate ~ fixed.dat$PctPublicCoverage,
        main = "Public Insurance Coverage vs Cancer Rate",
        xlab = "% Public Coverage",
        ylab = "Cancer Rate")
```



Income, however, is not strongly correlated with the cancer rate.

```
boxplot(fixed.dat$cancerRate ~ fixed.dat$medIncome, main = "Median Income vs Cancer Rate", xlab = "Media
```

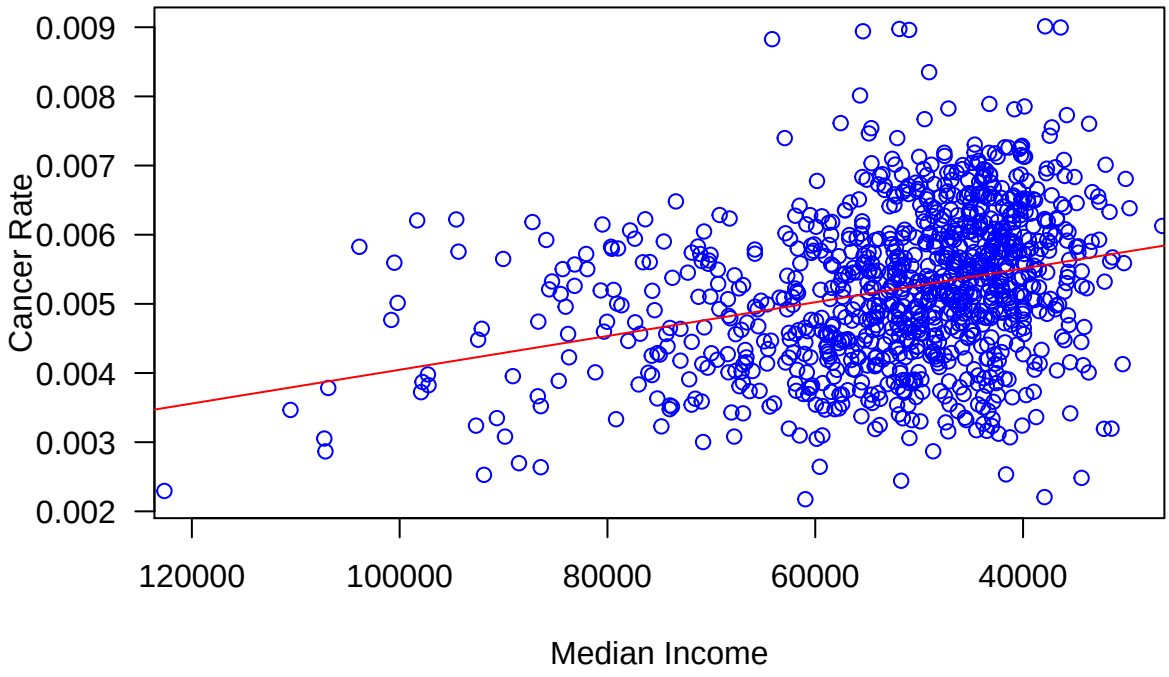


Lower medIncome and higher `PctPublicCoverage` have the most impact on `cancerRate`. Public insurance often has limited coverage for cancer treatment. Neither private insurance coverage nor poverty seem to be correlated with `cancerRate`.

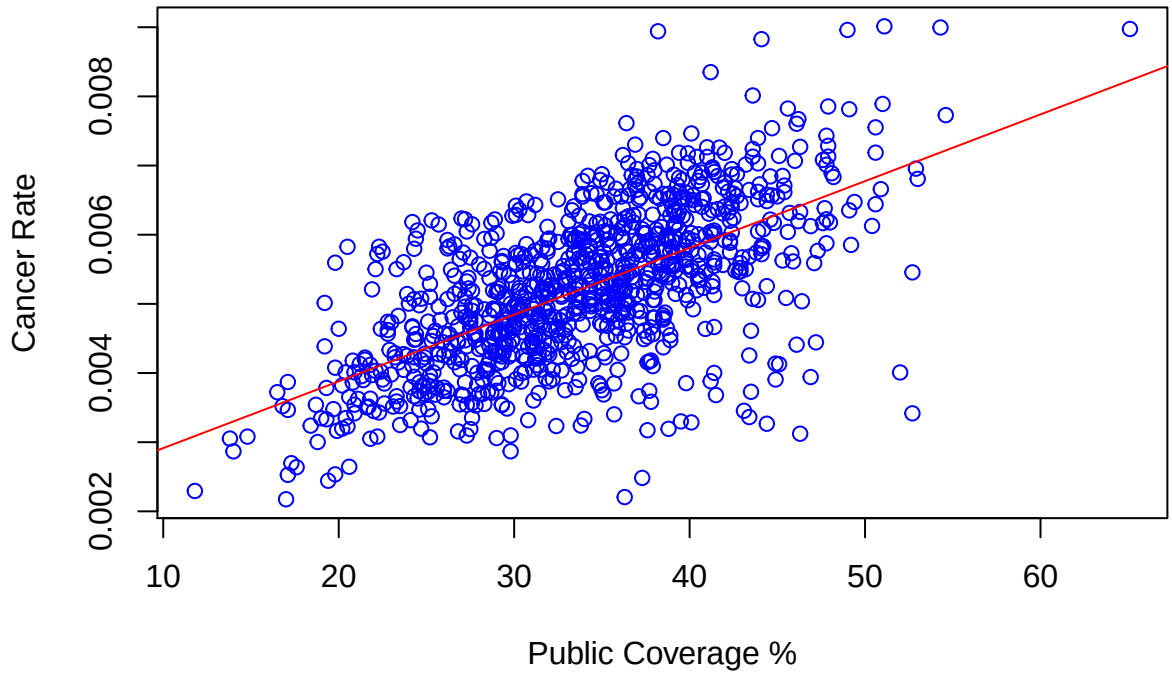
```
highInc.columns = c("avgAnnCount", "cancerRate", "Geography", "medIncome",
                    "PctPublicCoverage", "PctPrivateCoverage", "povertyPercent",
                    "AvgHouseholdSize", "deathToCancerRate")
highInc.dat = fixed.dat[fixed.dat$avgAnnCount > 250, highInc.columns]
nrow(highInc.dat)
```

```
## [1] 981
```

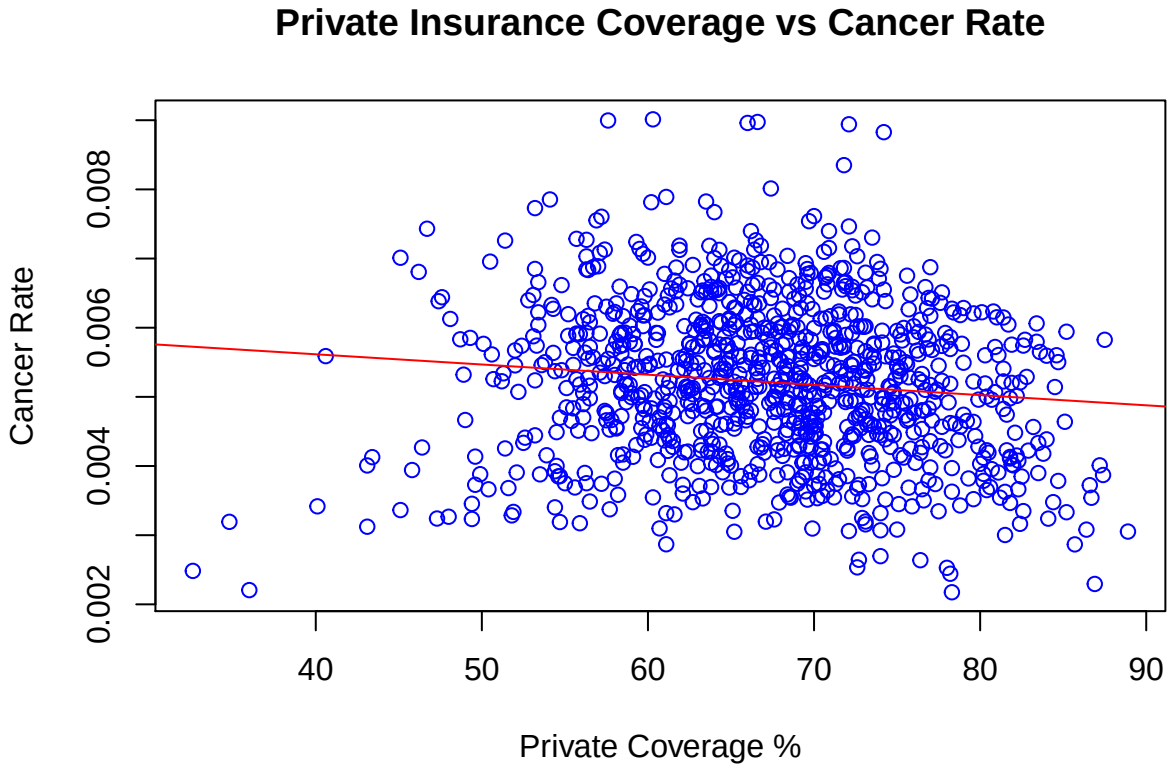
```
plot(highInc.dat$cancerRate ~ highInc.dat$medIncome,
     main = "Median Income to Cancer Rate",
     xlab = "Median Income",
     ylab = "Cancer Rate",
     las = 1,
     xlim = c(120000, 30000),
     col = 4)
abline(lm(highInc.dat$cancerRate ~ highInc.dat$medIncome), col = 2)
```



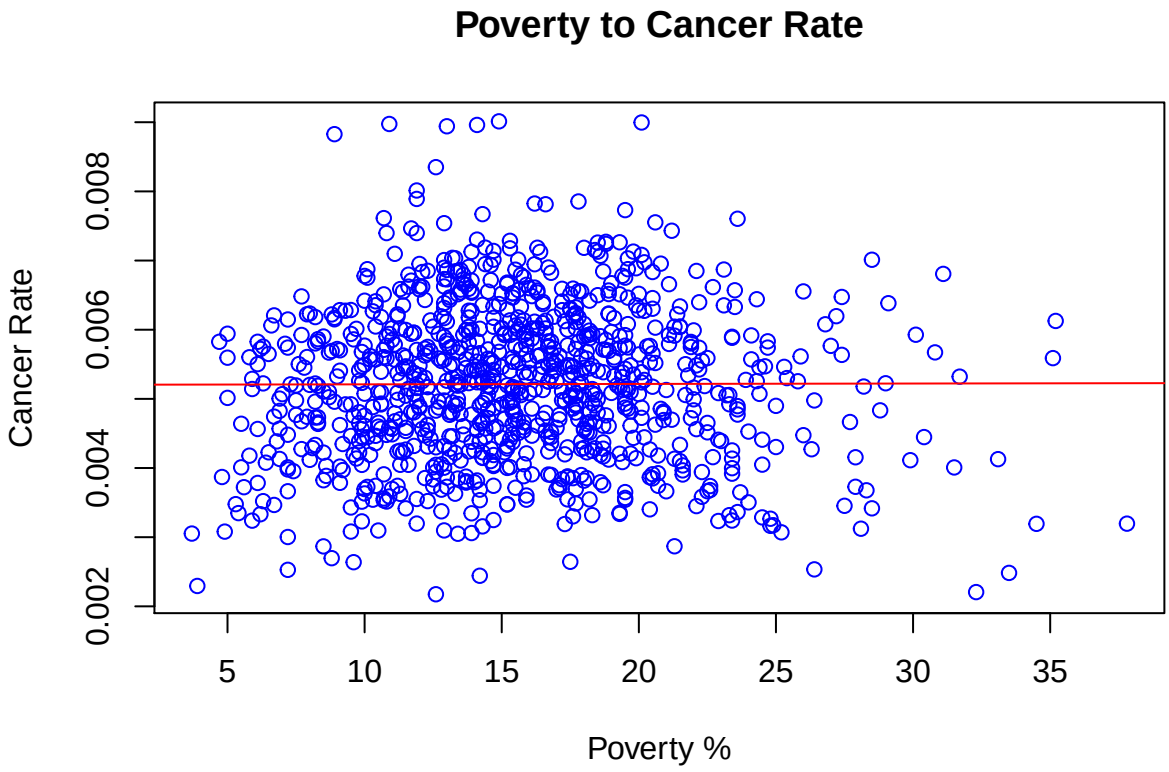

```
plot(highInc.dat$cancerRate ~ highInc.dat$PctPublicCoverage,
     main = "Public Insurance Coverage vs Cancer Rate",
     xlab = "Public Coverage %",
     ylab = "Cancer Rate",
     col = 4)
abline(lm(highInc.dat$cancerRate ~ highInc.dat$PctPublicCoverage), col = 2)
```




```
plot(highInc.dat$cancerRate ~ highInc.dat$PctPrivateCoverage,
     main = "Private Insurance Coverage vs Cancer Rate",
     xlab = "Private Coverage %",
     ylab = "Cancer Rate",
     col = 4)
abline(lm(highInc.dat$cancerRate ~ highInc.dat$PctPrivateCoverage), col = 2)
```

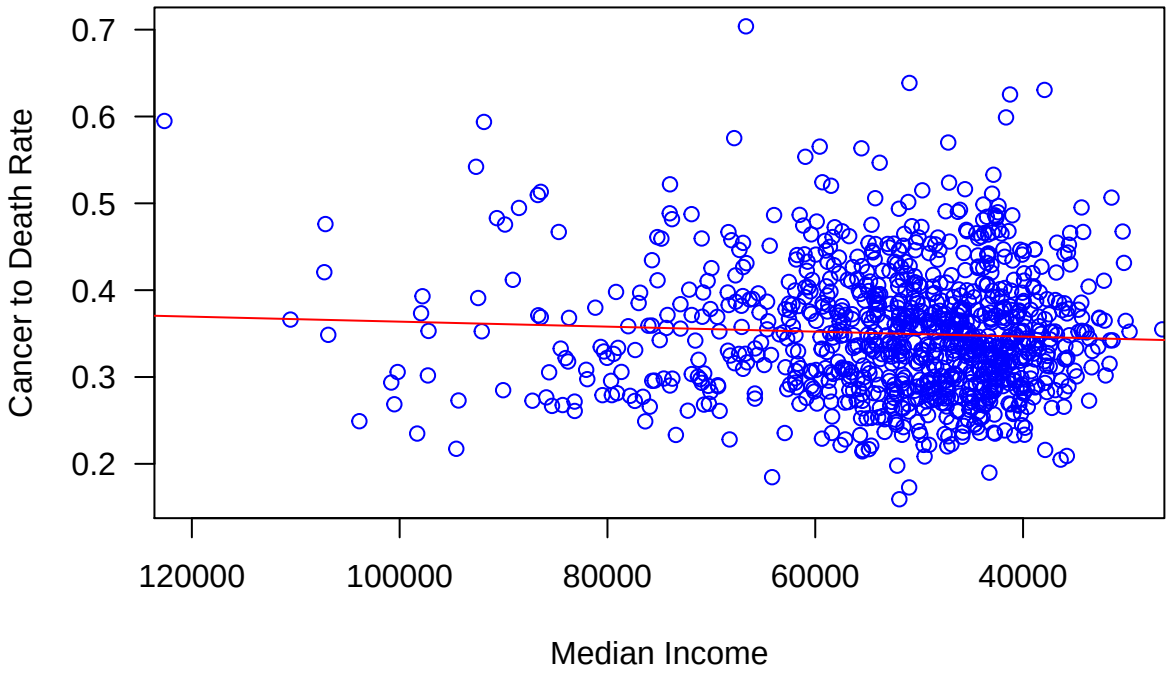



```
plot(highInc.dat$cancerRate ~ highInc.dat$povertyPercent,
     main = "Poverty to Cancer Rate",
     xlab = "Poverty %",
     ylab = "Cancer Rate",
     col = 4)
abline(lm(highInc.dat$cancerRate ~ highInc.dat$povertyPercent), col = 2)
```

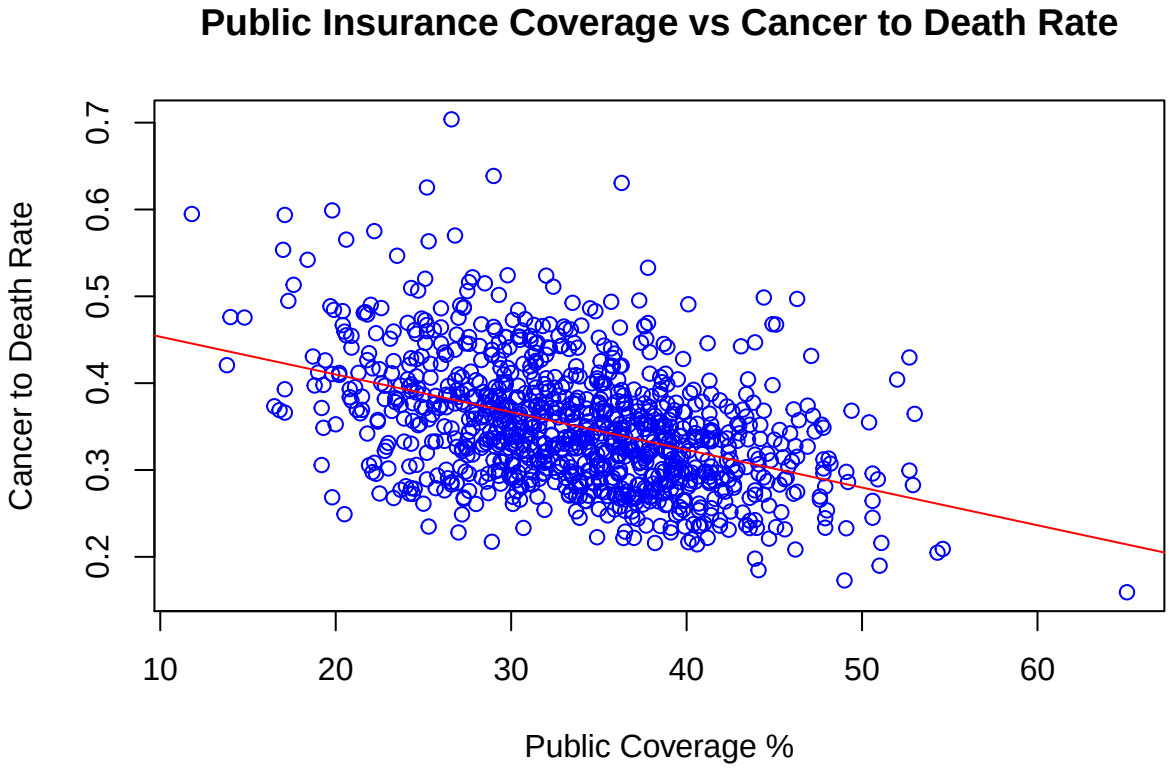


Now let's examine deathToCancerRate with medIncome and PctPublicCoverage. The results are less correlated compare to the previous plots. Since the data are not a direct cancer motality, the results suggest there are other factors to death rate, but `medIncome` and `PctPublicCoverage` are still correlated with cancer incident per county.

```
plot(highInc.dat$deathToCancerRate ~ highInc.dat$medIncome,
     main = "Median Income to Cancer to Death Rate",
     xlab = "Median Income",
     ylab = "Cancer to Death Rate",
     las = 1,
     xlim = c(120000, 30000),
     col = 4)
abline(lm(highInc.dat$deathToCancerRate ~ highInc.dat$medIncome), col = 2)
```




```
plot(highInc.dat$deathToCancerRate ~ highInc.dat$PctPublicCoverage,
     main = "Public Insurance Coverage vs Cancer to Death Rate",
     xlab = "Public Coverage %",
     ylab = "Cancer to Death Rate",
     col = 4)
abline(lm(highInc.dat$deathToCancerRate ~ highInc.dat$PctPublicCoverage), col = 2)
```



Race

We will now look at the race based fields `PctWhite`, `PctBlack`, `PctAsian`, `PctOtherRace`.

`PctWhite` and `PctBlack` are very highly correlated (over 0.8) and as we saw before, all race related fields are mutually exhaustive. Based on correlation coefficents, none of the race fields have a strong positive or negative linear correlation (>0.5) with the cancer mortality rates.

As can be seen below, relationship of a field with cancerRate and with deathToCancerRate are reverse. For example, PctWhite has -0.37 correlation with deathtocancerrate but 0.27 with cancer rate. While this might make one think that counties with higher white population are more likely to get cancer but less likely to have a cancer mortality, it might be a play of numbers here, and because of how deathToCancerRate is calculated (we have cancerRate in the denominator and a very weakly correlated deathRate in the numerator)

```
cor(fixed.dat[ , c("cancerRate","PctWhite","PctBlack","PctAsian", "PctOtherRace")], use = "complete.obs
```

```
##              cancerRate   PctWhite    PctBlack    PctAsian PctOtherRace
## cancerRate    1.0000000  0.2697025 -0.07522330 -0.27313964  -0.33509361
## PctWhite      0.2697025  1.0000000 -0.82884014 -0.26002025  -0.22597290
## PctBlack     -0.0752233 -0.8288401  1.00000000  0.01004758  -0.02696288
## PctAsian     -0.2731396 -0.2600203  0.01004758  1.00000000   0.20179716
## PctOtherRace -0.3350936 -0.2259729 -0.02696288  0.20179716   1.00000000
```

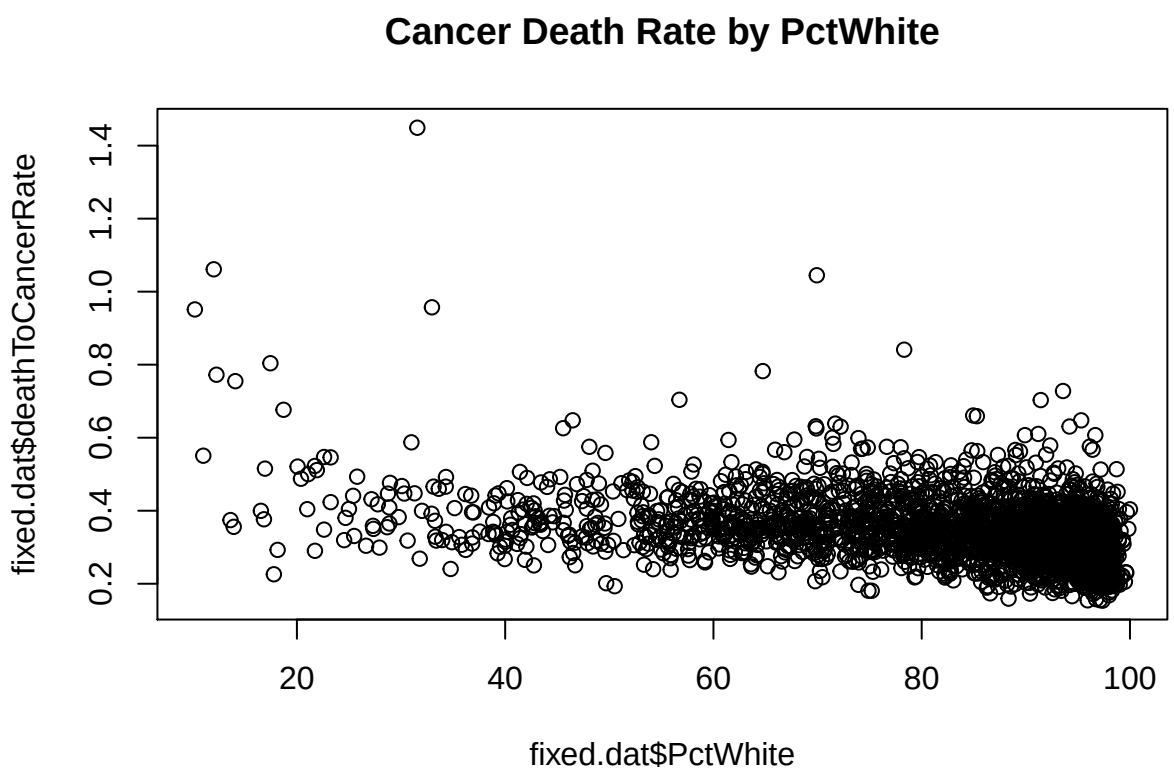
```
cor(fixed.dat[ , c("deathToCancerRate","PctWhite","PctBlack","PctAsian", "PctOtherRace")], use = "compl
```

```
##                   deathToCancerRate   PctWhite    PctBlack    PctAsian
## deathToCancerRate         1.0000000 -0.3752686  0.19060555  0.18575286
## PctWhite                 -0.3752686  1.0000000 -0.82884014 -0.26002025
## PctBlack                  0.1906056 -0.8288401  1.00000000  0.01004758
## PctAsian                  0.1857529 -0.2600203  0.01004758  1.00000000
## PctOtherRace              0.2017591 -0.2259729 -0.02696288  0.20179716
```

```
##                    PctOtherRace
## deathToCancerRate   0.20175907
## PctWhite           -0.22597290
## PctBlack           -0.02696288
## PctAsian            0.20179716
## PctOtherRace        1.00000000
```

Since PctWhite is the dominant field we can begin by looking at this field in more detail. We can see that predominant data has a higher PctWhite population. Since we see higher cancer rates at lower values of PctWhite, we could explore that further.

```
plot(fixed.dat$PctWhite,fixed.dat$deathToCancerRate,main="Cancer Death Rate by PctWhite")
```



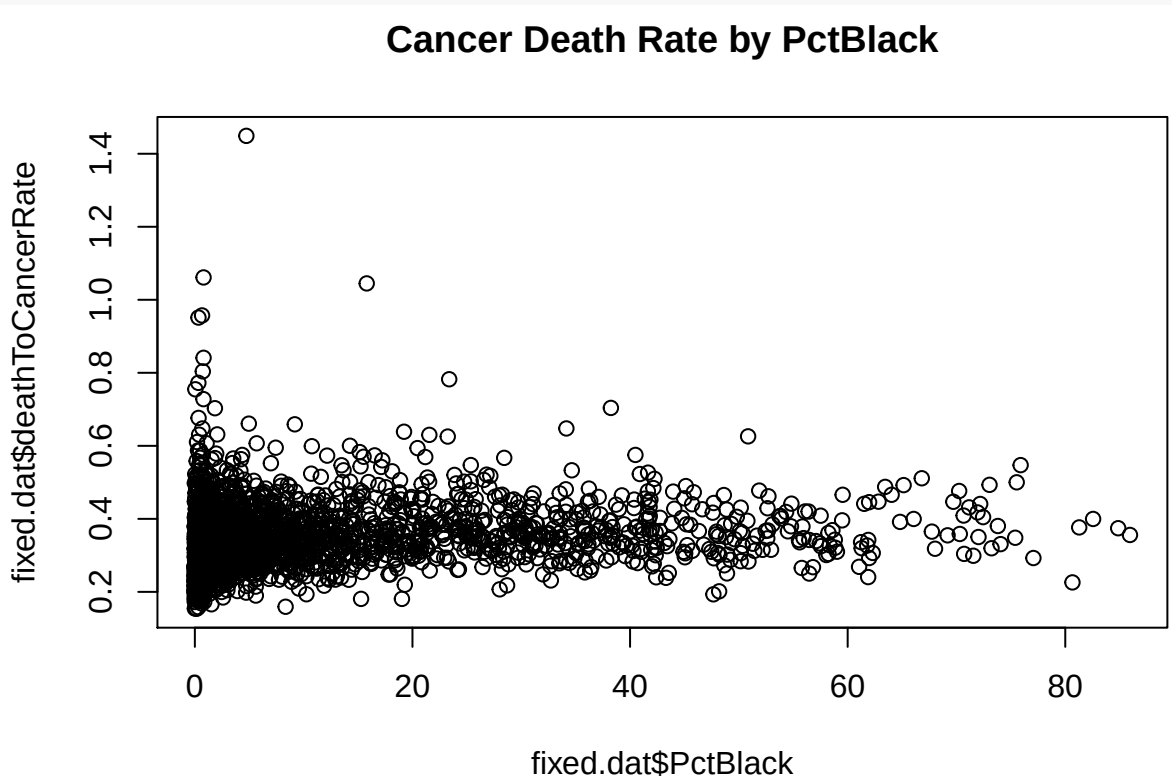
Let us create an indicator to segment populations based on pct of white population and compare the mean cancer mortality rates. We can see populations with less than 40pct whites have a higher cancer mortality rate. We do not know yet if these are statistically significant differences, but that is outside the scope of this EDA analysis.

```
fixed.dat$PctWhite_Segment<-"40 or more PctWhite"
fixed.dat$PctWhite_Segment[fixed.dat$PctWhite<40]<-"Less than 40 PctWhite"
aggregate(fixed.dat$deathToCancerRate,list(fixed.dat$PctWhite_Segment),mean)
```

```
##                 Group.1         x
## 1   40 or more PctWhite 0.3375629
## 2 Less than 40 PctWhite 0.4387967
```

Let us check PctBlack against cancer mortality rate.

```
plot(fixed.dat$PctBlack,fixed.dat$deathToCancerRate,main="Cancer Death Rate by PctBlack")
```



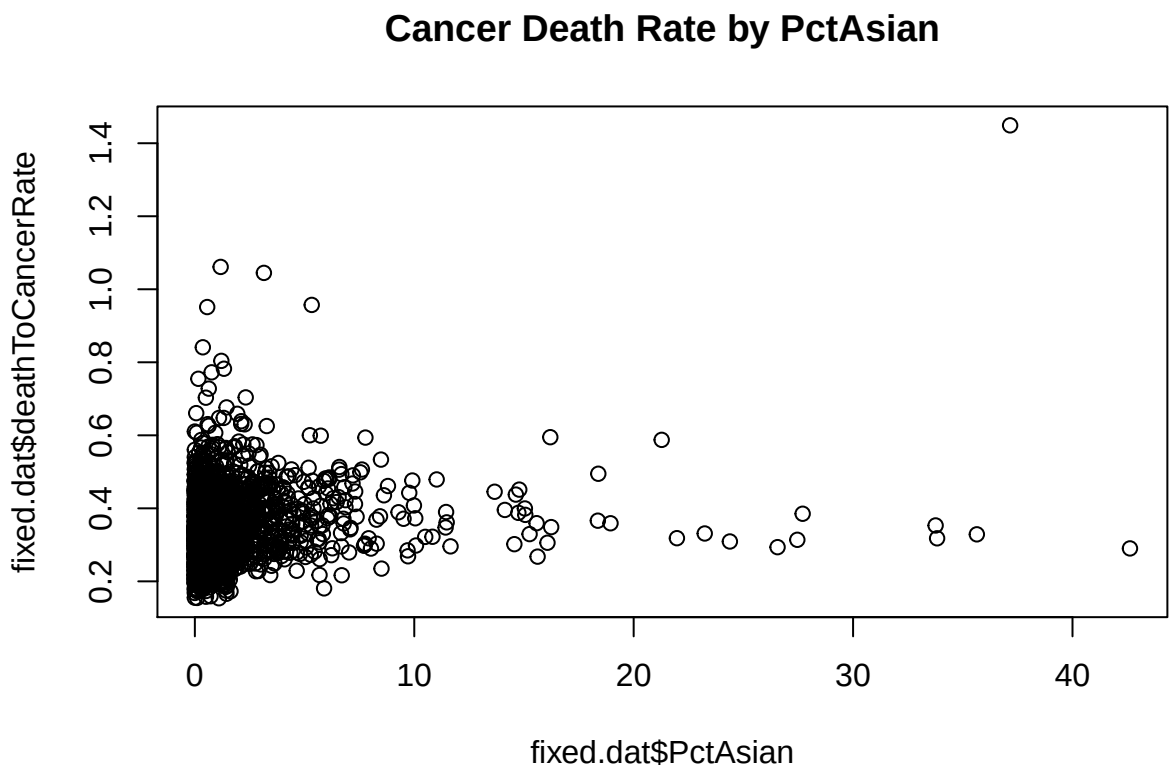
We can see most populations have a low rate of PctBlack. Segmenting by a rough threhsold of 15 PctBlack, we see that counties with more than 20pct black have a slightly higher cancer mortality rate, but the difference is not very high (without testing for statistical significance or running any other population comparison tests)

```
fixed.dat$PctBlack_Segment<-"15 or more PctBlack"
fixed.dat$PctBlack_Segment[fixed.dat$PctBlack<15]<-"Less than 15 PctBlack"
aggregate(fixed.dat$deathToCancerRate,list(fixed.dat$PctBlack_Segment),mean)
```

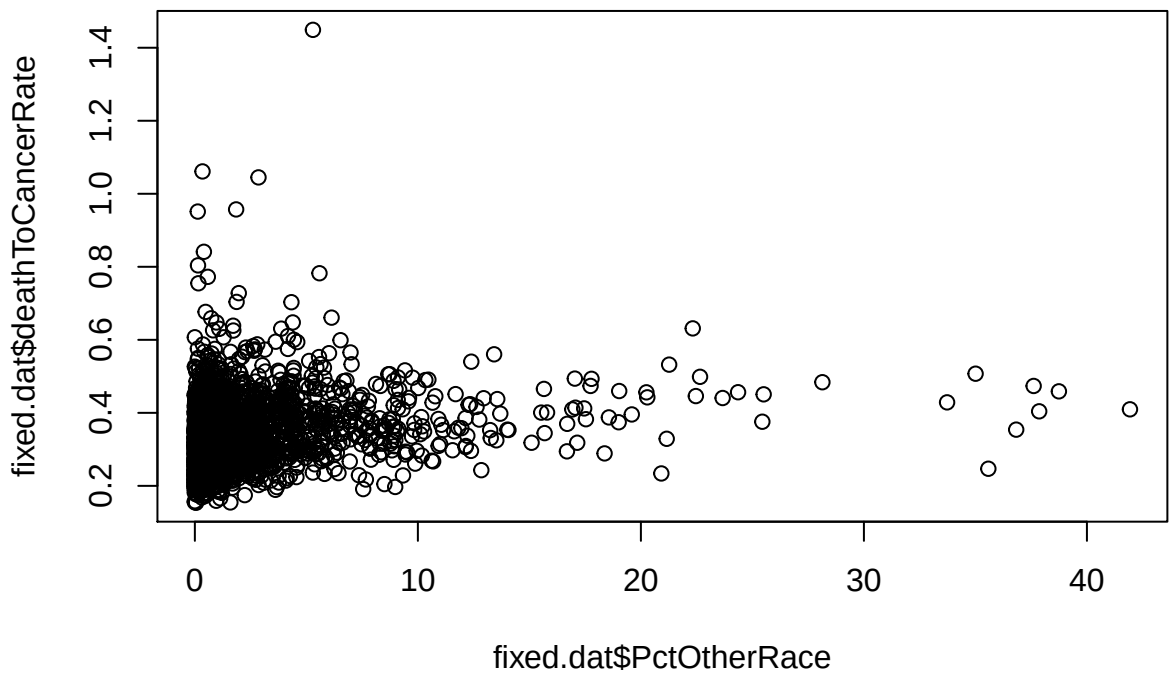
```
##                 Group.1         x
## 1   15 or more PctBlack 0.3736745
## 2 Less than 15 PctBlack 0.3317199
```

PctAsian and PctOtherRace do not seem to have any noticeable pattern with cancer mortality rate.

```
plot(fixed.dat$PctAsian,fixed.dat$deathToCancerRate,main="Cancer Death Rate by PctAsian")
```



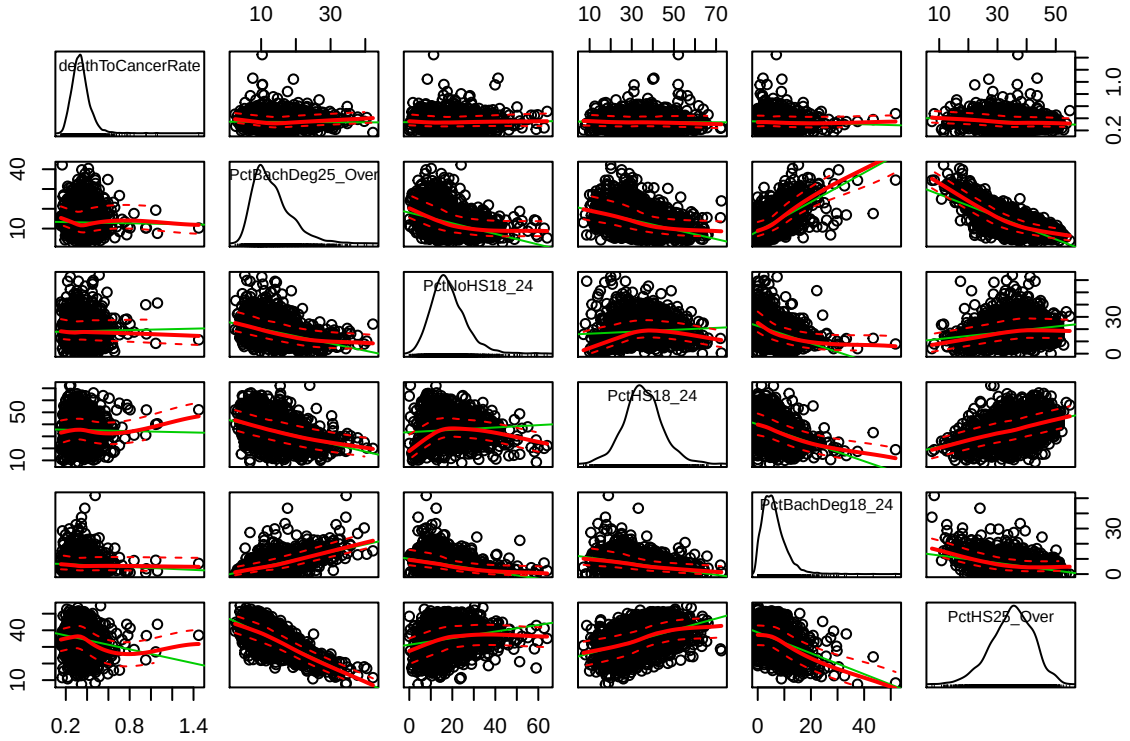

```
plot(fixed.dat$PctOtherRace,fixed.dat$deathToCancerRate,main="Cancer Death Rate by PctOtherRace")
```



Education and Employment

We will now cover the variables related to education and employment.It does not seem like education or employment has any key role in influencing cancer death rates.

```
edu_fields<-c("PctBachDeg25_Over","PctNoHS18_24","PctNoHS18_24","PctHS18_24","PctBachDeg18_24","PctHS25_
scatterplotMatrix(~ deathToCancerRate + PctBachDeg25_Over + PctNoHS18_24 + PctHS18_24 + PctBachDeg18_24
```




```
round(cor(fixed.dat$deathToCancerRate,fixed.dat[,edu_fields],use="pairwise.complete.obs"),3)
```

```
##      PctBachDeg25_Over PctNoHS18_24 PctNoHS18_24.1 PctHS18_24
## [1,]            -0.016        0.022          0.022     -0.019
##      PctBachDeg18_24 PctHS25_Over PctEmployed16_Over PctUnemployed16_Over
## [1,]          -0.061        -0.17               0.006                0.222
```

Intuitively, one would imagine a higher education would mean better employment, income, better insurance coverage and access to more expensive treatment and thus lower cancer mortality rate, but the data does not seem to capture that. We can see that the PctBachDeg25_Over has high correlation to medIncome, PctPrivateCoverage and PctPrivateCoverage indicating that more educated people have better and private coverage, but this does not translate to lower cancer mortality rates as per the data.

```
round(cor(fixed.dat$medIncome,fixed.dat[,edu_fields],use="pairwise.complete.obs"),3)
```

```
##      PctBachDeg25_Over PctNoHS18_24 PctNoHS18_24.1 PctHS18_24
## [1,]             0.711       -0.305         -0.305      -0.19
##      PctBachDeg18_24 PctHS25_Over PctEmployed16_Over PctUnemployed16_Over
## [1,]           0.512       -0.472               0.699               -0.457
```

```
round(cor(fixed.dat$PctPrivateCoverage,fixed.dat[,edu_fields],use="pairwise.complete.obs"),3)
```

```
##      PctBachDeg25_Over PctNoHS18_24 PctNoHS18_24.1 PctHS18_24
## [1,]             0.601        -0.47          -0.47     -0.235
##      PctBachDeg18_24 PctHS25_Over PctEmployed16_Over PctUnemployed16_Over
```

```
## [1,]           0.496       -0.214              0.689               -0.62
```

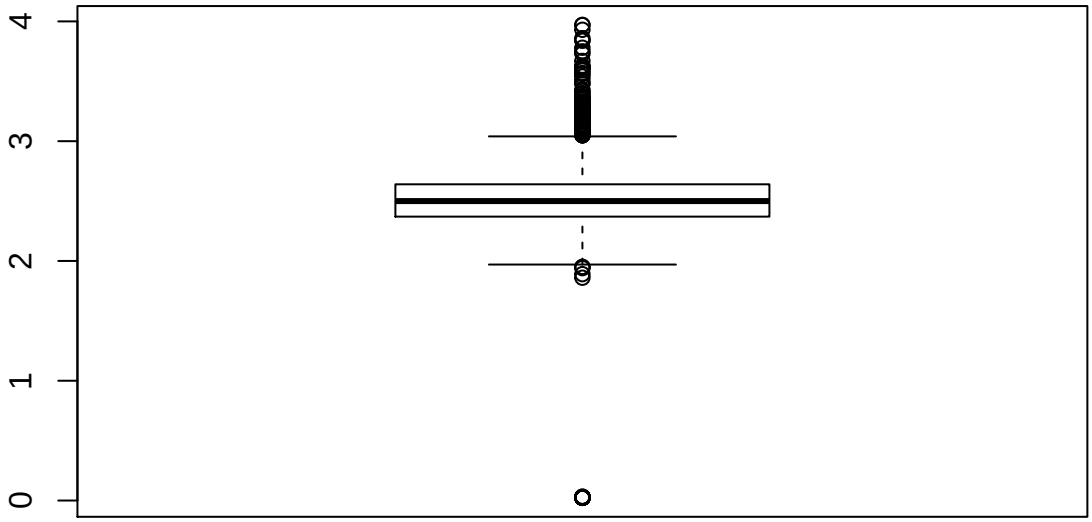
```
round(cor(fixed.dat$PctEmpPrivCoverage,fixed.dat[,edu_fields],use="pairwise.complete.obs"),3)
```

```
##      PctBachDeg25_Over PctNoHS18_24 PctNoHS18_24.1 PctHS18_24
## [1,]             0.533       -0.445         -0.445     -0.238
##      PctBachDeg18_24 PctHS25_Over PctEmployed16_Over PctUnemployed16_Over
## [1,]            0.47       -0.208              0.703               -0.476
```

Family Structure & Age

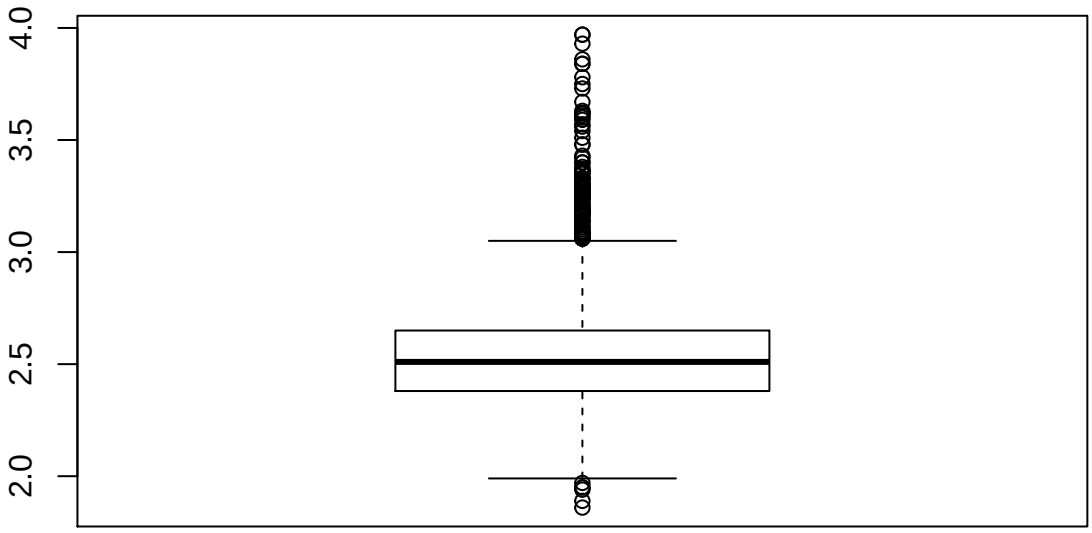
There are two three variables related to family structure: `AvgHouseholdSize`, `PercentMarried`, `PctMarriedHouseholds`.

```
# Look for outliers
boxplot(fixed.dat$AvgHouseholdSize)
```

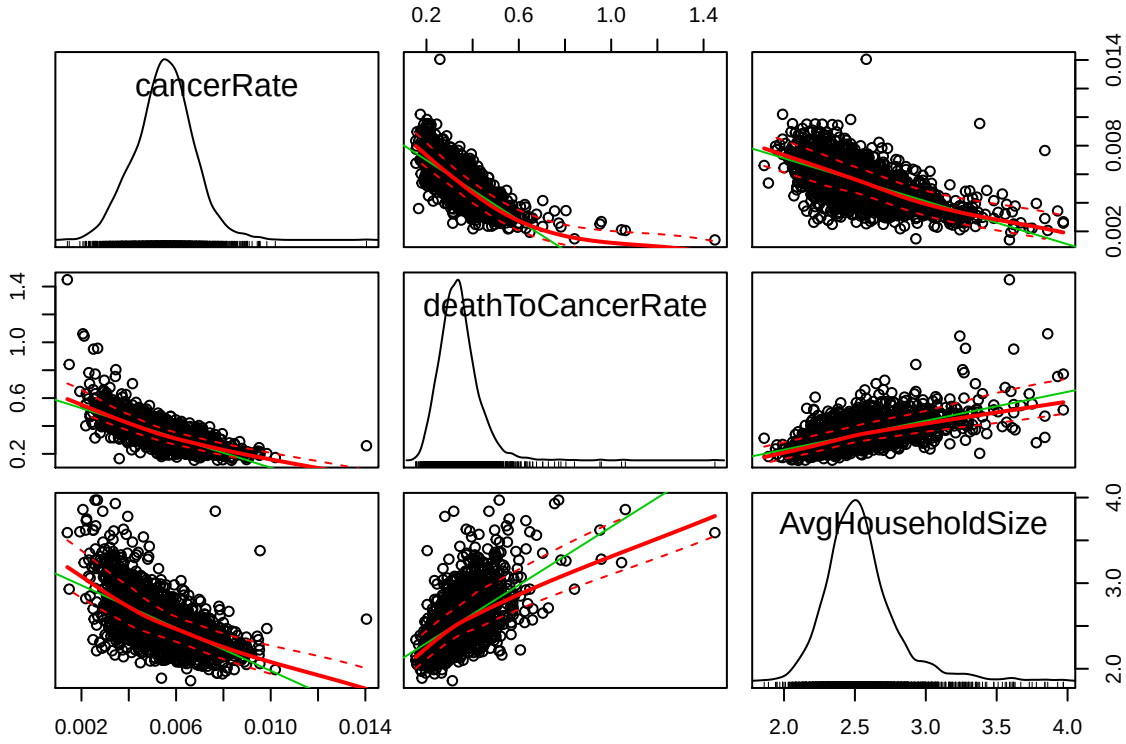


Average household size is close 0 for 57 observations, but really these numbers shouldn't be less than 1. Instead of making some assumption about what the reasoning is for these numbers, we will opt to remove them from this particular set of analyses.

```
household.dat <- subset(fixed.dat, AvgHouseholdSize >= 1)
boxplot(household.dat$AvgHouseholdSize)
```




```
scatterplotMatrix(~ cancerRate + deathToCancerRate + AvgHouseholdSize,
                  data = household.dat)
```



It appears that both cancer rates decrease and death to cancer rates increase with increases in household size.

```
cor(household.dat$cancerRate, household.dat$AvgHouseholdSize)
```

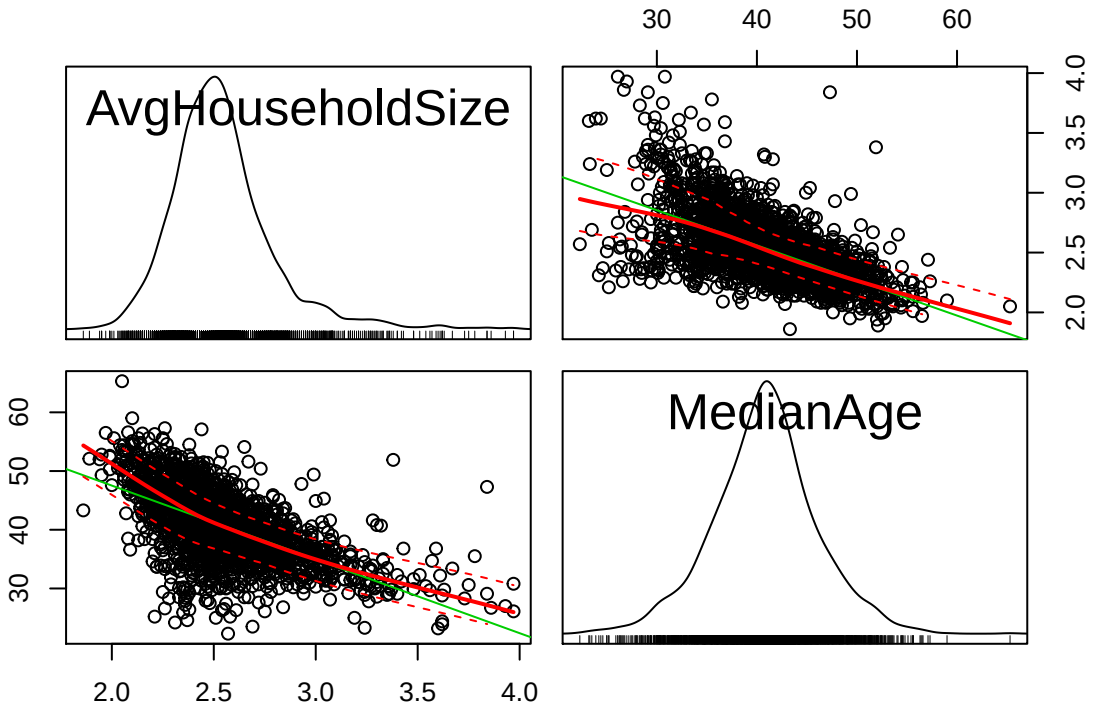
```
## [1] -0.613028
```

```
cor(household.dat$deathToCancerRate, household.dat$AvgHouseholdSize)
```

```
## [1] 0.5940166
```

What does this mean? It *could* mean that there are fewer incidents of cancer in larger households. This is most likely because larger households on average means younger people. Let's test that assumption.

```
scatterplotMatrix(~ AvgHouseholdSize + MedianAge, data = household.dat)
```



This is reasonably well correlated:

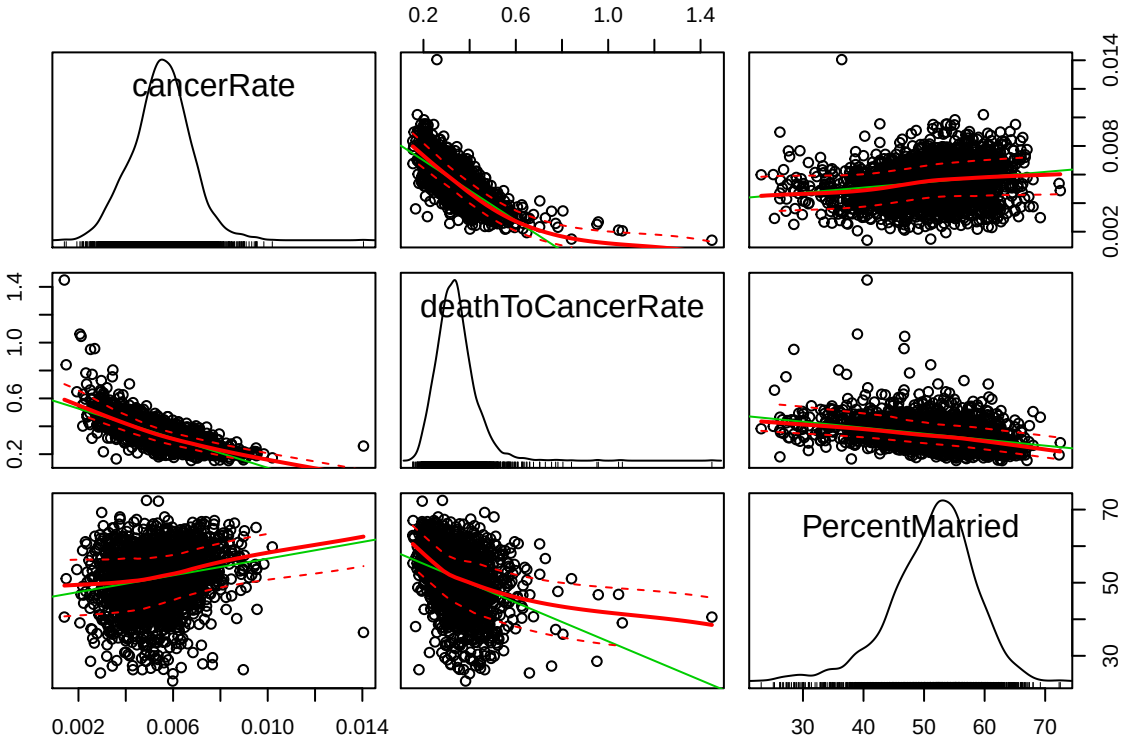
```
cor(household.dat$AvgHouseholdSize, household.dat$MedianAge)
```

```
## [1] -0.6072252
```

Based on that correlation, it can't be said that household size alone contributes to lower rates of cancer. Interestingly, the ratio of death to cancer increases with larger household sizes. What could explain this? It could be that there are more deaths unrelated to cancer with younger groups. This seems feasible, since younger groups of people are less likely to have routine, easily detectable cancer.

What about marriage?

```
scatterplotMatrix(~ cancerRate + deathToCancerRate + PercentMarried, data = fixed.dat)
```




```
cor(fixed.dat$cancerRate, fixed.dat$PercentMarried)
```

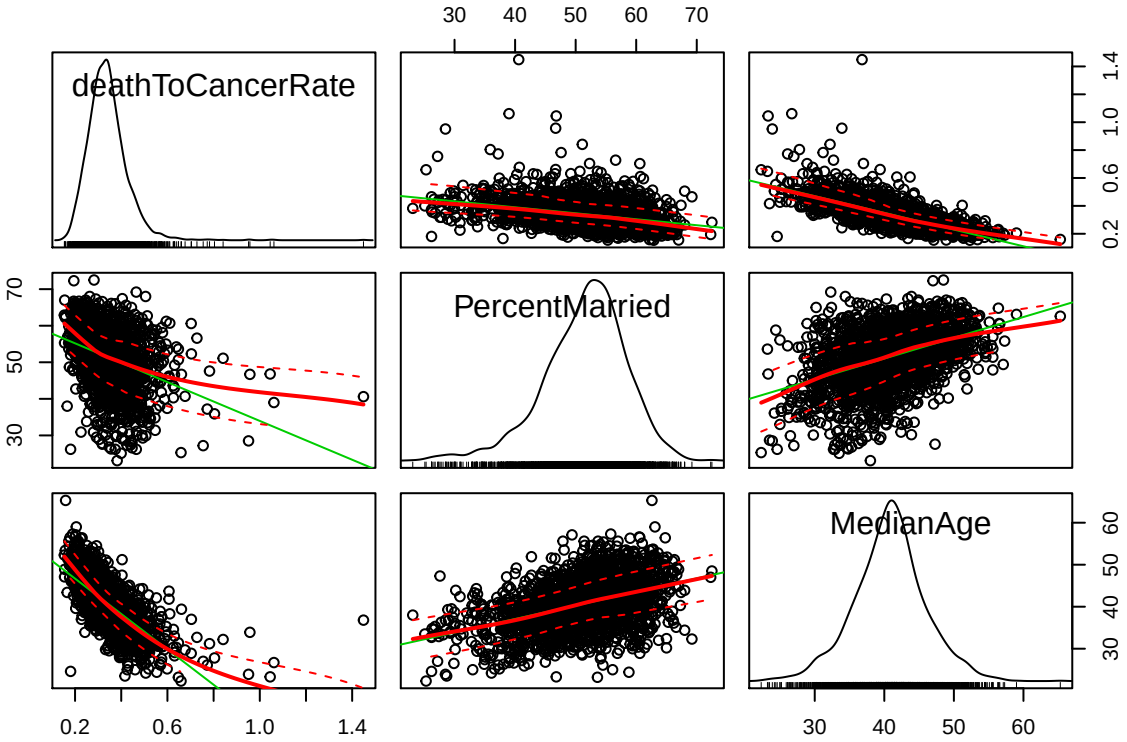
```
## [1] 0.2043757
```

```
cor(fixed.dat$deathToCancerRate, fixed.dat$PercentMarried)
```

```
## [1] -0.337221
```

The correlation between percent married and cancer rate is only ~0.2. There may be something here, but it's pretty weak. The correllation between the death to cancer rate and percent married is slightly higher, at -0.34. This could mean that there is a slight tendency for a lower incidence of cancer mortality for married couples. This could possibly be explained by earlier detection, stronger motivation in the form of spousal urging to take early and strong action to conquer the cancer, etc. Could there be any confounding factors, such as the age of the population?

```
scatterplotMatrix(~ deathToCancerRate + PercentMarried + MedianAge, data = fixed.dat)
```



It appears that median age percent married are positively correlated, more so than death to cancer rate and percent married.

```
cor(fixed.dat$PercentMarried, fixed.dat$MedianAge)
```

```
## [1] 0.4278646
```

Again, as ages increase, the percentage of married people increases, so the two are dependent, and cancer mortality rates can't necessarily be attributed to marriage or age alone.

Analysis of Secondary Effects

There are several variables that may have confounding effects, positive and negative. Below are a few examples of fields in the data and fields that are not in the data.

- Income on insurance coverage rates
- Marital status on size of household
- Insurance on treatment related data
- Gender on several factors from income to mortality rates (due to differences in types of cancer)
- Type of cancer
- Sun or radiation exposure on age and geography (indirectly)

Conclusion

In conclusion, this analysis has attempted to explore per-county cancer data to look at characteristics that may have impact on cancer mortality rates. The data provided has some minor housekeeping issues, but those issues were mostly handled easily. The biggest problem with the data given our task at hand is that there is no way to find an exact cancer mortality rate. The only semblance of such a value was a deaths-to-cancer-incidents percentage. This percentage is most likely very misleading, as there are much more prevalent causes of death in many communities. The second biggest issue is that the eventual objective

of this assigment is to look for "social interventions" that could be implemented, but there are many more data points that could be useful in terms of possible social interventions, assuming correlation.

Given the above pitfalls of the data, from a preliminary EDA analysis on county level averages, there do appear to be correlations between the death-to-cancer rate in a county to things such as age, insurance, and race. Factors such as education and household status do not seem to have strong correlations.

Communities with the following characteristics (one or more) can be targeted for interventions:

- Counties with higher rates of public insurance and lower median income
- Counties with less than a 40% white population have higher cancer rates
- Counties with higher rate of smaller households and unmarried people

Based on the strengths of correlation coefficients and differences in average cancer mortality rates, focusing on counties with high rates of public insurance and less than 40% white population are likely to have a larger impact.