Exploratory Data Analysis - Cancer Mortality Rates

W203 Lab Project (Fall 2018)

Authors: Lina Gurevich, Duda Espindola, Jonathan D'Souza

Executive Summary

Given a Data Set for cancer incidences for a select group of counties.... this study attempts to explore the relationships between the outcome variable : Death Rate and other key independent variables. After some exploration and discussion, we decided to focus in on variables related to Health insurance and understand their immpact (if any) on cancer mortality. Furthermore we studied how variables related to income level and cancer incident rate interacted (both with each other and with mortality rate). Our conclusions are summarized at the end of this brief.

Detailed Steps and Findings

Initial Loading and Validation of Data Set

Set Up

```
raw_data<-read.csv("cancer.csv")  #Assumes file in current working directory
cancer.df<-raw_data #Keep one copy of raw data as is
```

Summarize Data Set

```
str(cancer.df)
```

```
## 'data.frame':    3047 obs. of  30 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ avgAnnCount         : num  1397 173 102 427 57 ...
##  $ medIncome           : int  61898 48127 49348 44243 49955 52313 37782 40189 42579 60397 ...
##  $ popEst2015          : int  260131 43269 21026 75882 10321 61023 41516 20848 13088 843954 ...
##  $ povertyPercent      : num  11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
##  $ binnedInc           : Factor w/ 10 levels "(34218.1, 37413.8]",..: 9 6 6 4 6 7 2 2 3 8 ...
##  $ MedianAge           : num  39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
##  $ MedianAgeMale       : num  36.9 32.2 44 42.2 47.8 43.5 42.2 50.8 48.4 34.7 ...
##  $ MedianAgeFemale     : num  41.7 33.7 45.8 43.4 48.9 48 43.5 52.5 49.8 37 ...
##  $ Geography           : Factor w/ 3047 levels "Abbeville County, South Carolina",..: 1459 1460 1464
##  $ AvgHouseholdSize    : num  2.54 2.34 2.62 2.52 2.34 2.58 2.42 2.24 2.38 2.65 ...
##  $ PercentMarried      : num  52.5 44.5 54.2 52.7 57.8 50.4 54.1 52.7 55.9 50 ...
##  $ PctNoHS18_24        : num  11.5 6.1 24 20.2 14.9 29.9 26.1 27.3 34.7 15.6 ...
##  $ PctHS18_24          : num  39.5 22.4 36.6 41.2 43 35.1 41.4 33.9 39.4 36.3 ...
##  $ PctSomeCol18_24     : num  42.1 64 NA 36.1 40 NA NA 36.5 NA NA ...
##  $ PctBachDeg18_24     : num  6.9 7.5 9.5 2.5 2 4.5 5.8 2.2 1.4 7.1 ...
##  $ PctHS25_Over        : num  23.2 26 29 31.6 33.4 30.4 29.8 31.6 32.2 28.8 ...
##  $ PctBachDeg25_Over   : num  19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
```

```
##  $ PctEmployed16_Over  : num  51.9 55.9 45.9 48.3 48.2 44.1 51.8 40.9 39.5 56.6 ...
##  $ PctUnemployed16_Over: num  8 7.8 7 12.1 4.8 12.9 8.9 8.9 10.3 9.2 ...
##  $ PctPrivateCoverage  : num  75.1 70.2 63.7 58.4 61.6 60 49.5 55.8 55.5 69.9 ...
##  $ PctEmpPrivCoverage  : num  41.6 43.6 34.9 35 35.1 32.6 28.3 25.9 29.9 44.4 ...
##  $ PctPublicCoverage   : num  32.9 31.1 42.1 45.3 44 43.2 46.4 50.9 48.1 31.4 ...
##  $ PctWhite            : num  81.8 89.2 90.9 91.7 94.1 ...
##  $ PctBlack            : num  2.595 0.969 0.74 0.783 0.27 ...
##  $ PctAsian            : num  4.822 2.246 0.466 1.161 0.666 ...
##  $ PctOtherRace        : num  1.843 3.741 2.747 1.363 0.492 ...
##  $ PctMarriedHouseholds: num  52.9 45.4 54.4 51 54 ...
##  $ BirthRate           : num  6.12 4.33 3.73 4.6 6.8 ...
##  $ deathRate           : num  165 161 175 195 144 ...
```

The data set has data that spans 30 variables for 3047 different counties (based on the number of levels in the Geography variable being the same as total number of observations). We note that most of the variables are numeric variables, with the exception of Geography and Binned Income which are categorical.

Validation and cleaning of variables.

Check for NA Values

```
colSums(is.na(cancer.df))
```

```
##                    X          avgAnnCount            medIncome
##                    0                    0                    0
##           popEst2015       povertyPercent            binnedInc
##                    0                    0                    0
##            MedianAge        MedianAgeMale      MedianAgeFemale
##                    0                    0                    0
##            Geography     AvgHouseholdSize       PercentMarried
##                    0                    0                    0
##         PctNoHS18_24           PctHS18_24      PctSomeCol18_24
##                    0                    0                 2285
##      PctBachDeg18_24         PctHS25_Over    PctBachDeg25_Over
##                    0                    0                    0
##   PctEmployed16_Over PctUnemployed16_Over   PctPrivateCoverage
##                  152                    0                    0
##   PctEmpPrivCoverage    PctPublicCoverage             PctWhite
##                    0                    0                    0
##             PctBlack             PctAsian         PctOtherRace
##                    0                    0                    0
## PctMarriedHouseholds            BirthRate            deathRate
##                    0                    0                    0
```

There are 2 variables with null values: PctSomeCol18_24 and PctEmployed16_Over.

Clean up of MedianAge variable From the summary of the Median Age it is clear that there are some outliers above 100 years given the max of 624 compared to median & mean in the 40s.

```
#Check medianAge based on summary
summary(cancer.df$MedianAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.30   37.70   41.00   45.27   44.00  624.00
```

Looking at just the outliers, they are clearly erroneous values.

```
#Check medianAge based on summary

ageoutliers<-cancer.df[cancer.df$MedianAge>100,]
summary(ageoutliers$MedianAge) #
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   349.2   461.1   499.2   492.6   522.3   624.0
```

Based on the order of magnitude difference (around 10), we assume that there was a data capture error and divide all these values by 10 to create a normalized data set.

```
#Divide outliers by 10
cancer.df$MedianAge[cancer.df$MedianAge>150]<-cancer.df$MedianAge/10  # Set outlier values to NA
summary(cancer.df$MedianAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.42   37.60   40.90   40.46   43.80   65.30
```

Validation & Clean up of avgAnnCount

Annual Incident Count is better expressed as a percentage of county population.

```
cancer.df$AnnCountPercent<-100*cancer.df$avgAnnCount/cancer.df$popEst2015
summary(cancer.df$AnnCountPercent)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.09281   0.48020   0.56240   2.32400   0.64870 236.80000
```

Having more than an incident count of more than 100% is clearly not possible (more incidents of cancer diagnoses than the population of the county). We look for where the outliers may be coming from.

```
#Assuming anything over 50% incident rate has to be an error
outliers<-cancer.df[cancer.df$AnnCountPercent>50,]
summary(outliers$avgAnnCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1963    1963    1963    1963    1963    1963
```

It is clear that all these observations have the exact same erroneous value for Average Annual Count. We will set these to NULL and recalculate average annual incident count as a percent of population.

```
error_value<-outliers[1,"avgAnnCount"]
#Assuming any observation with this value is an error, set them to NA
cancer.df$avgAnnCount[cancer.df$avgAnnCount==error_value]<-NA
#Recalculate percentages
cancer.df$AnnCountPercent<-with(cancer.df,100*avgAnnCount/popEst2015)
summary(cancer.df$AnnCountPercent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.1403  0.4747  0.5532  0.5507  0.6283  1.4050     206
```

Data Transformation for Analysis

We're going to explore a set of variables that represent the levels of health insurance coverage for individual counties.There are three variables in the original dataset that are related to insurance:

Table 1: Primary variables for exploration

Variable Name	Description
PctPrivateCoverage	Percentage of the population with private insurance coverage
PctPublicCoverage	Percentage of the population with public insurance coverage
PctEmpPrivCoverage	Percentage of the population with employer-sponsored private insurance coverage

For the purposes of our explanatory analysis, we would like to conduct a more comprehensive research on various types and levels of insurance coverage and their effects on the mortality rates, so it makes sense to define a few more variables that can be derived from the original dataset. For example, we would like to include data about the populations with no insurance coverage, as well as the observations where individuals have both private and public insurance. It can also be more revealing to treat the employer-sponsored coverage as a relative proportion of the private coverage rather than an absolute value. We will compute these as both continuous and binned discrete variables (for easier analysis) as follows:

Table 2: Additional Derived Primary variables for exploration

Variable Name	Description
PctPNoCoverage	Percentage of the population with no insurance coverage
PctDoubleCoverage	Percentage of the population with both private and public insurance coverage
EmpSponsoredPct	Percentage of the private insurance sponsored by employers
PctPublicCoverageCat	Percentage of the population with public insurance coverage binned into 10 categories
PctPrivateCoverageCat	Percentage of the population with private insurance coverage binned into 10 categories
PctPublicCoverageCat	Percentage of the population with employer sponsored private insurance binned into 10 categories
IncomeCat	Median Income binned

We will now add these new variables to our original dataset:

```
cancer.df$PctDoubleCoverage=cancer.df$PctPublicCoverage + cancer.df$PctPrivateCoverage - 100
cancer.df$PctDoubleCoverage[cancer.df$PctDoubleCoverage < 0] = 0
cancer.df$PctNoCoverage = 100 - cancer.df$PctPublicCoverage - cancer.df$PctPrivateCoverage
cancer.df$PctNoCoverage[cancer.df$PctNoCoverage < 0] = 0
cancer.df$EmpSponsoredPct = cancer.df$PctEmpPrivCoverage / cancer.df$PctPrivateCoverage * 100
cancer.df$PctPublicCoverageCat<-cut(cancer.df$PctPublicCoverage, seq(0,100,10), right=FALSE)
cancer.df$PctPrivateCoverageCat<-cut(cancer.df$PctPrivateCoverage, seq(0,100,10), right=FALSE)
cancer.df$PctEmpPrivCoverageCat<-cut(cancer.df$PctEmpPrivCoverage, seq(0,100,10), right=FALSE)
cancer.df$IncomeCat<-cut(cancer.df$medIncome, seq(0,160000,20000), right=FALSE, labels=c("0 - 20k","20k
```

Our key variables in this investigation will be deathRate (target variable) and several indpendent variables representing insurance coverage for counties' populations.

Cancer Mortality Rate (deathRate variable)

Let's start with the target variable and summarize it:

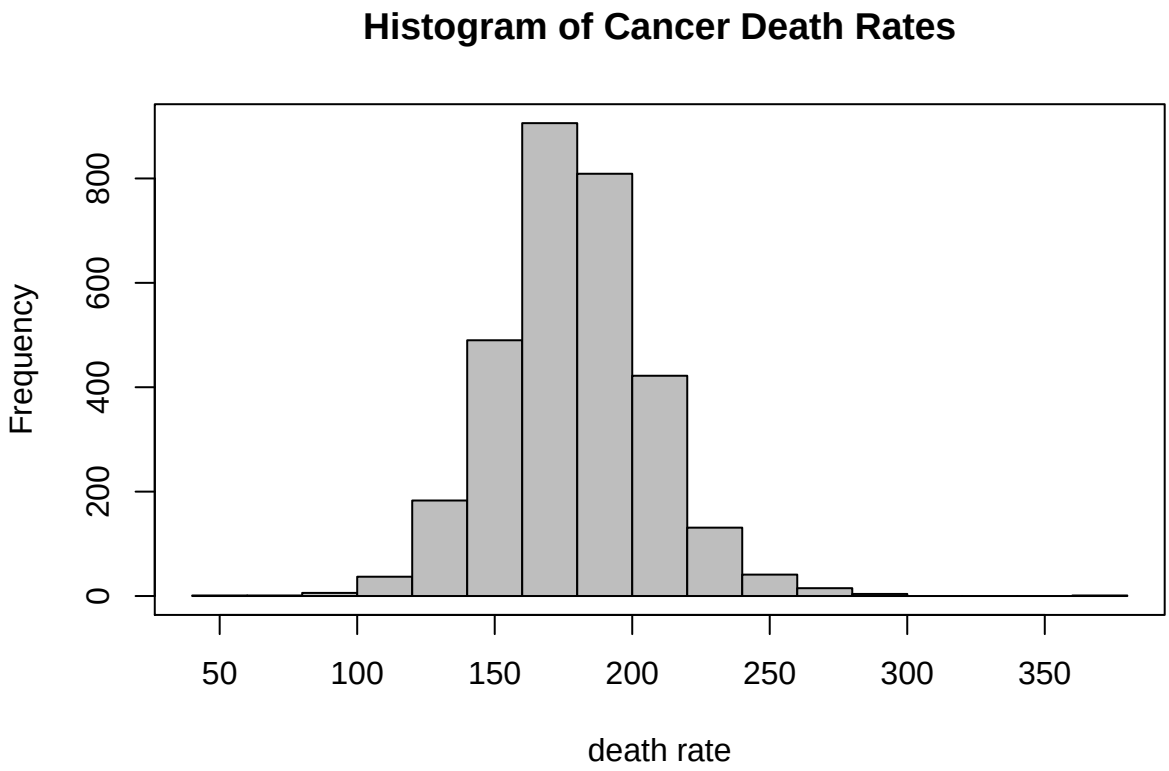
```
summary(cancer.df$deathRate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    59.7   161.2   178.1   178.7   195.2   362.8
```

We see that this is a metric variable with its mean and median values very close to each other. There are no missing values and no obviously wrong or suspicious outliers.

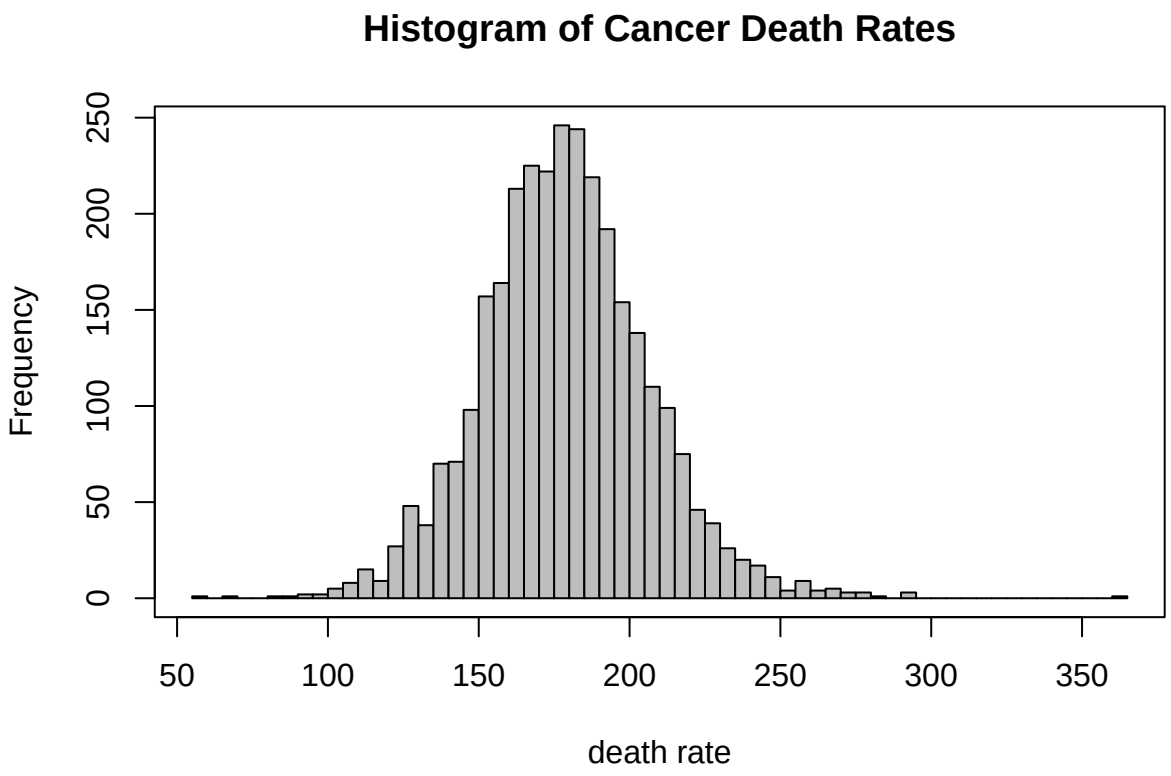
To better visualize the variable's values distribution, we plot the histogram.

```
with(cancer.df, hist(deathRate,  col = "gray",
                     main="Histogram of Cancer Death Rates",
                     xlab="death rate"))
box()
```



As we can see from the output, the default method for selecting the number of bins produced too few bins, which might obscure some interesting features in the data. A better result is achieved by setting the binning rule to the one proposed by Freedman and Diaconis. Fortunately, hist() function has a built-in option for this:

```
with(cancer.df, hist(deathRate,  breaks='FD', col = "gray",
                     main="Histogram of Cancer Death Rates",
                     xlab="death rate"))
box()
```



Now we have a much higher level of detail and can easily infer that deathRate variable distribution is very close to the normal one, with a notable outliers on the far right of the histogram.

Let's explore the extreme outliers with deathRate over 300 and see if we can find anything unusual in these observations. To find out how many outliers are there, we'll use the nrow() function:

```
nrow(cancer.df[cancer.df$deathRate > 300,])
```

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

Turns out there's only one observation with this property, so let's examine it a bit closer.

```
str(cancer.df[cancer.df$deathRate > 300,])
```

```
## 'data.frame':    1 obs. of  38 variables:
##  $ X                    : int 1490
##  $ avgAnnCount          : num 214
##  $ medIncome            : int 40207
##  $ popEst2015           : int 15234
##  $ povertyPercent       : num 24.3
##  $ binnedInc            : Factor w/ 10 levels "(34218.1, 37413.8]",..: 2
##  $ MedianAge            : num 40.3
##  $ MedianAgeMale        : num 42.3
##  $ MedianAgeFemale      : num 36.9
##  $ Geography            : Factor w/ 3047 levels "Abbeville County, South Carolina",..: 2762
##  $ AvgHouseholdSize     : num 2.58
##  $ PercentMarried       : num 36.4
##  $ PctNoHS18_24         : num 27
##  $ PctHS18_24           : num 45.1
##  $ PctSomeCol18_24      : num NA
##  $ PctBachDeg18_24      : num 0
##  $ PctHS25_Over         : num 37.4
```

```
##  $ PctBachDeg25_Over    : num 5.5
##  $ PctEmployed16_Over   : num NA
##  $ PctUnemployed16_Over : num 11.7
##  $ PctPrivateCoverage   : num 59.6
##  $ PctEmpPrivCoverage   : num 41
##  $ PctPublicCoverage    : num 35.8
##  $ PctWhite             : num 74
##  $ PctBlack             : num 21.6
##  $ PctAsian             : num 0.645
##  $ PctOtherRace         : num 1.53
##  $ PctMarriedHouseholds : num 50
##  $ BirthRate            : num 3.74
##  $ deathRate            : num 363
##  $ AnnCountPercent      : num 1.4
##  $ PctDoubleCoverage    : num 0
##  $ PctNoCoverage        : num 4.6
##  $ EmpSponsoredPct      : num 68.8
##  $ PctPublicCoverageCat : Factor w/ 10 levels "[0,10)","[10,20)",..: 4
##  $ PctPrivateCoverageCat: Factor w/ 10 levels "[0,10)","[10,20)",..: 6
##  $ PctEmpPrivCoverageCat: Factor w/ 10 levels "[0,10)","[10,20)",..: 5
##  $ IncomeCat            : Factor w/ 8 levels "0 - 20k","20k - 40k",..: 3
```

At first sight, nothing in the rest of the data stands out to provide a possible explanation for the high mortality rate (363). We might want to revisit this observation once we completed the rest of the analysis.

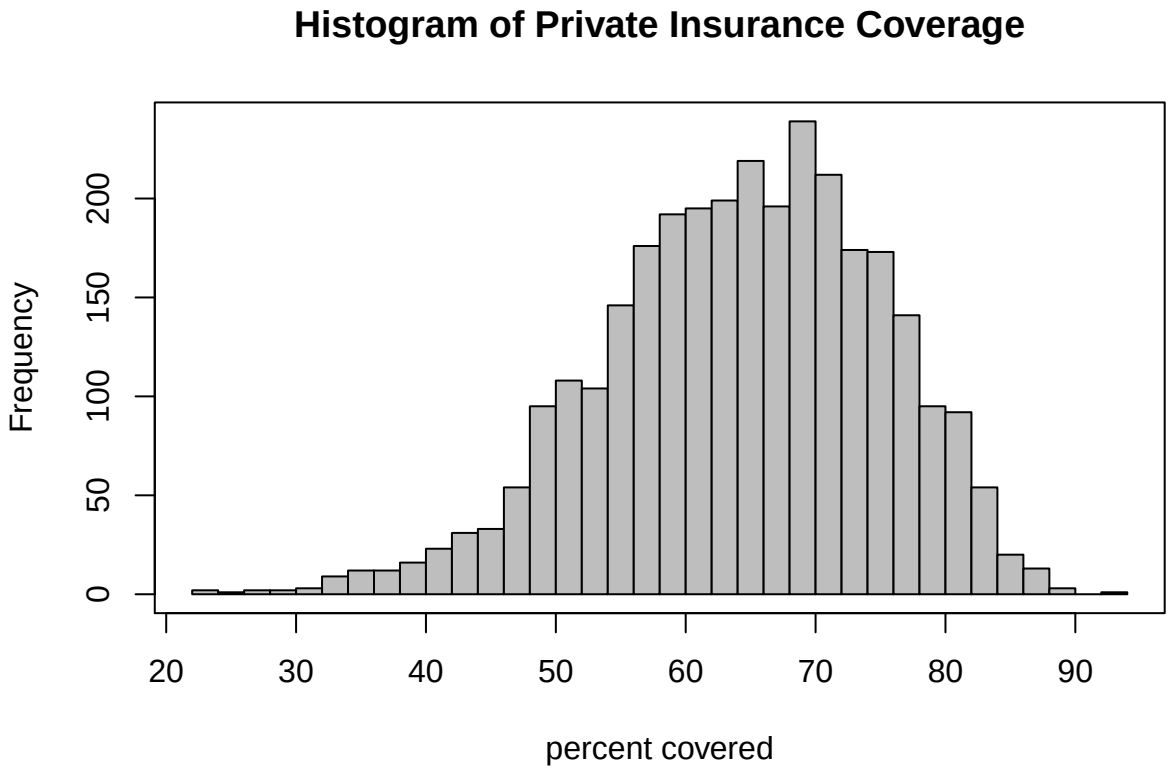
Private Insurance Coverage (PctPrivateCoverage variable)

Similar to our target variable, we summarize PctPrivateCoverage and generate its histogram:

```
summary(cancer.df$PctPrivateCoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.30   57.20   65.10   64.35   72.10   92.30
```

```
with(cancer.df, hist(PctPrivateCoverage, breaks="FD", col = "gray",
                     main="Histogram of Private Insurance Coverage",
                     xlab="percent covered"))
box()
```



We notice that the frequency distribution has some negative skew, with the majority of values falling between 55% and 75%. The data looks reasonable, with no obvious errors and missing values.

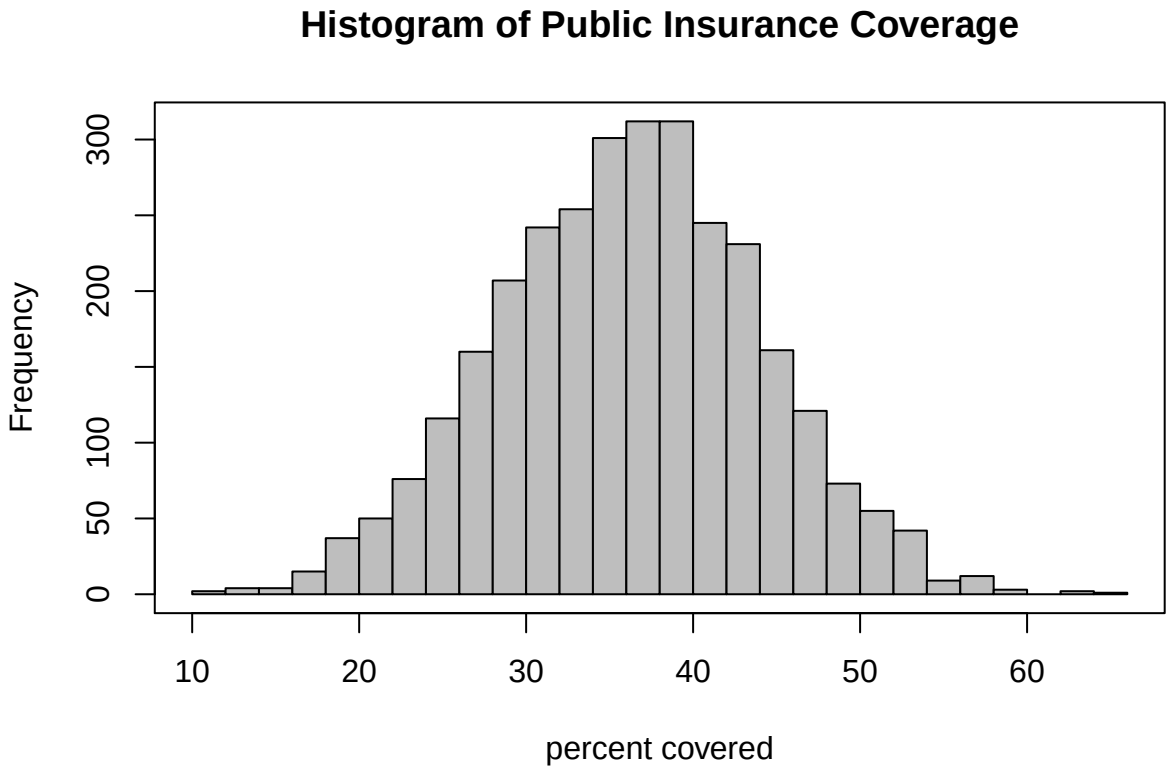
Public Insurance Coverage (PctPublicCoverage variable)

We repeat the steps executed above for the public insurance coverage:

```
summary(cancer.df$PctPublicCoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.20   30.90   36.30   36.25   41.55   65.10
```

```
with(cancer.df, hist(PctPublicCoverage, breaks="FD", col = "gray",
                     main="Histogram of Public Insurance Coverage",
                     xlab = "percent covered"))
box()
```



Compared to the private insurance coverage, the data is more evenly distributed and is much closer to the normal curve. The mean and median values are almost half of the ones for the private insurance coverage. From that we can infer that the private insurance is much more prevalent than the one sponsored by the state. Similar to PctPrivateCoverage, the public coverage variables doesn't show any obvious errors and there are no missing values.

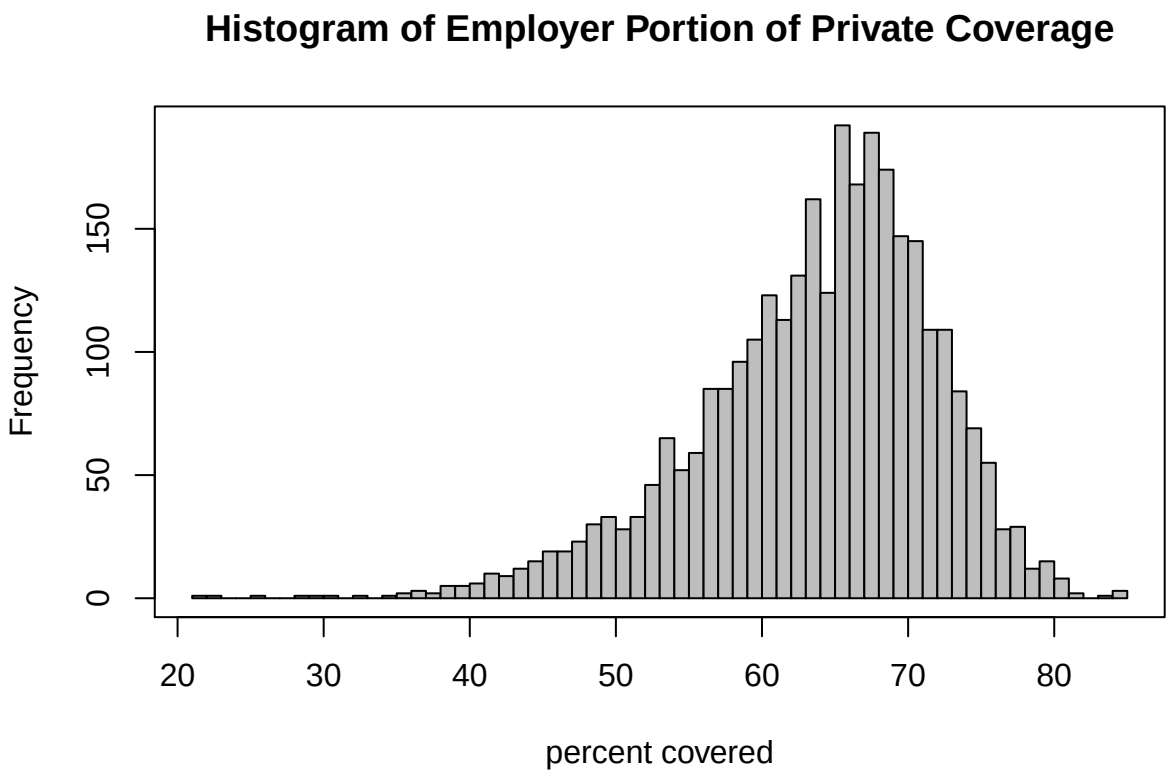
Employer-sponsored portion of the private coverage (EmpSponsoredPct variable)

After exploring the general category of the private coverage, we would like to examine what portion of the insurance are provided by employers:

```
summary(cancer.df$EmpSponsoredPct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.59   59.08   65.14   63.76   69.43   84.55
```

```
with(cancer.df, hist(EmpSponsoredPct, breaks="FD", col = "gray",
                     main="Histogram of Employer Portion of Private Coverage",
                     xlab = "percent covered"))
box()
```



The histogram tells us that employment is the major source of private insurance coverage in the counties: most of the values of EmpSponsoredPct variable fall between 60% and 70%.

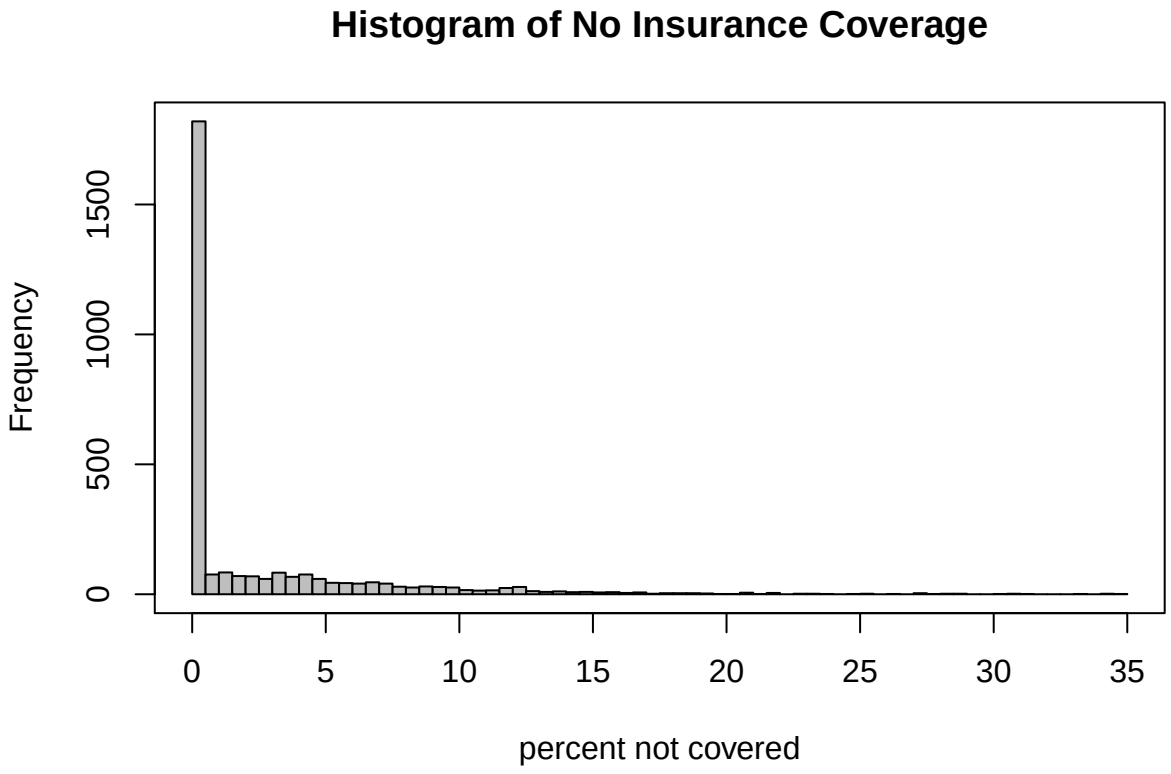
No insurance coverage (PctNoCoverage variable)

Let's summarize our generated variable that represents percentage of the population with no insurance coverage:

```
summary(cancer.df$PctNoCoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   2.595   3.750  34.600
```

```
with(cancer.df, hist(PctNoCoverage, breaks="FD", col = "gray",
                     main="Histogram of No Insurance Coverage",
                     xlab = "percent not covered"))
box()
```



Unlike the distributions we've seen so far, this variable has a major peak around 0, with the rest of the values tapering off in the shape of the long-tailed distribution. To get a better insight into the variable, we can generate the percentile metric:

```
quantile(cancer.df$PctNoCoverage, prob = seq(0, 1, length = 11), type = 5)
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##  0.0  0.0  0.0  0.0  0.0  0.0  0.6  2.7  4.9  8.7 34.6
```

The result shows that 80% of the observations have less than 5% of the population with no health insurance. We can safely infer then that the effect of this variable on the target will be minimal.

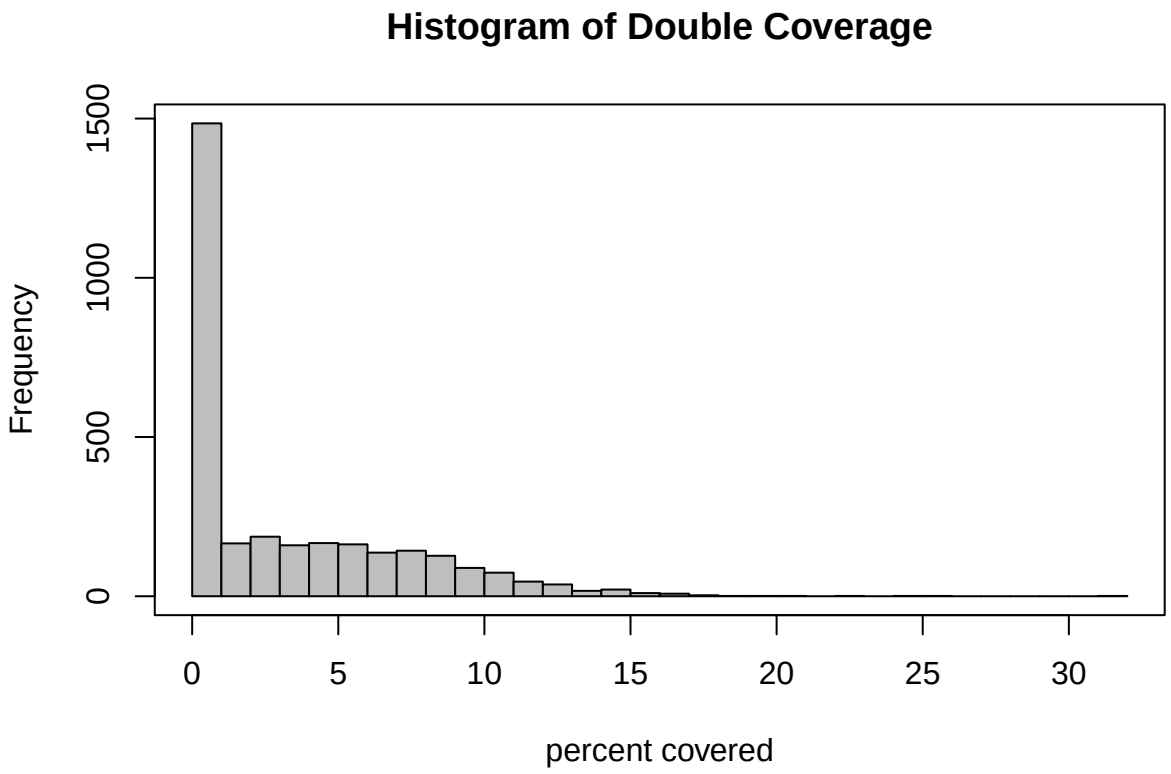
Coverage that includes both private and public components (PctDoubleCoverage variable)

We repeat the steps executed during the evaluation of PctNoCoverage variable:

```
summary(cancer.df$PctDoubleCoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.300   3.203   5.800  31.700
```

```
with(cancer.df, hist(PctDoubleCoverage, breaks="FD", col = "gray",
                     main="Histogram of Double Coverage",
                     xlab = "percent covered"))
box()
```




```
quantile(cancer.df$PctDoubleCoverage, prob = seq(0, 1, length = 11), type = 5)
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##  0.0  0.0  0.0  0.0  0.0  1.3  3.0  4.8  6.9  9.1 31.7
```

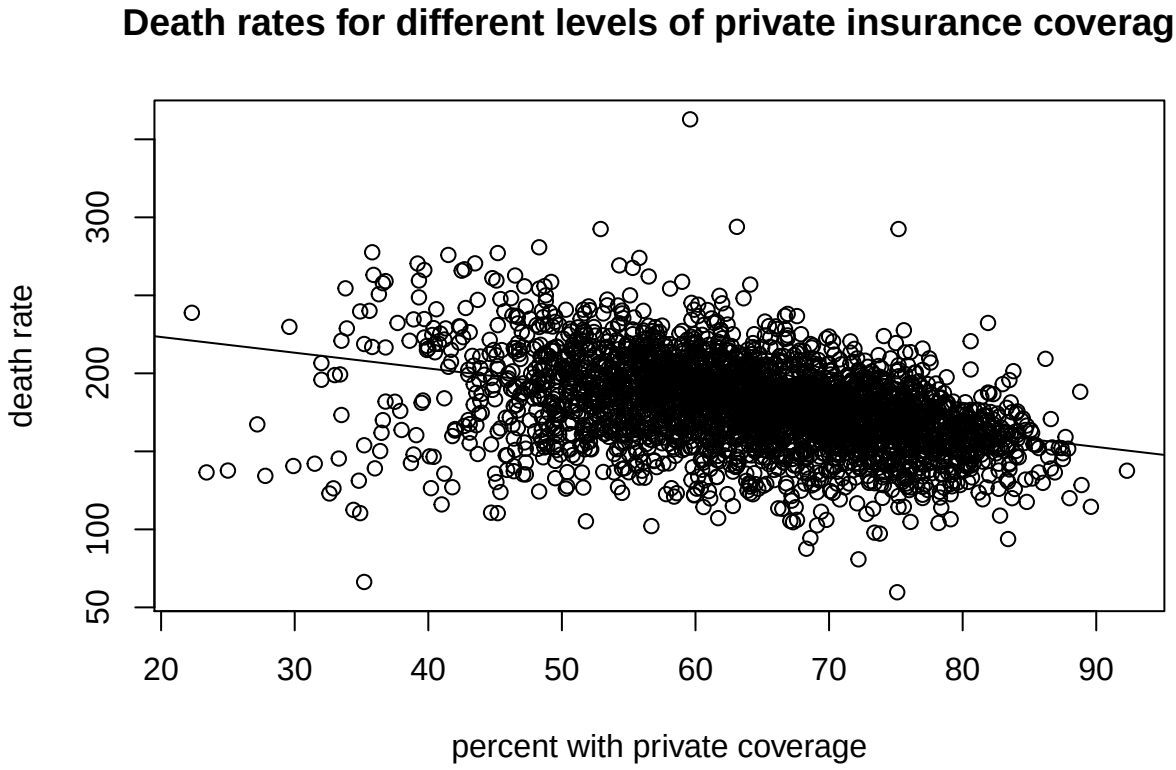
The result shows that 80% of the counties have less than 7% of the population with double health insurance. Therefore, similar to the previous case, its relative effect on the target variable will be minimal.

Analysis of Key Relationships

Mortality rates for different levels of private insurance coverage

Our first question is whether having access to a private insurance coverage is correlated with cancer mortality rates. A reasonable hypothesis would be that a cancer patient with a private insurance would be able to afford better treatment options. As a result, she or he will have better chances of survival, so we should expect negative correlation between deathRate and PctPrivateCoverage. Let's build a scatterplot showing the relationshoip between these two variables. In order to get a better insight into what linear relationship exists in the data, we add the ordinary least squares regression line to the plot and calculate the correlation.

```
plot(cancer.df$PctPrivateCoverage, cancer.df$deathRate,
     xlab = "percent with private coverage", ylab = "death rate",
     main = "Death rates for different levels of private insurance coverage")
abline(lm(cancer.df$deathRate ~ cancer.df$PctPrivateCoverage))
```




```
cor(cancer.df$deathRate, cancer.df$PctPrivateCoverage)
```

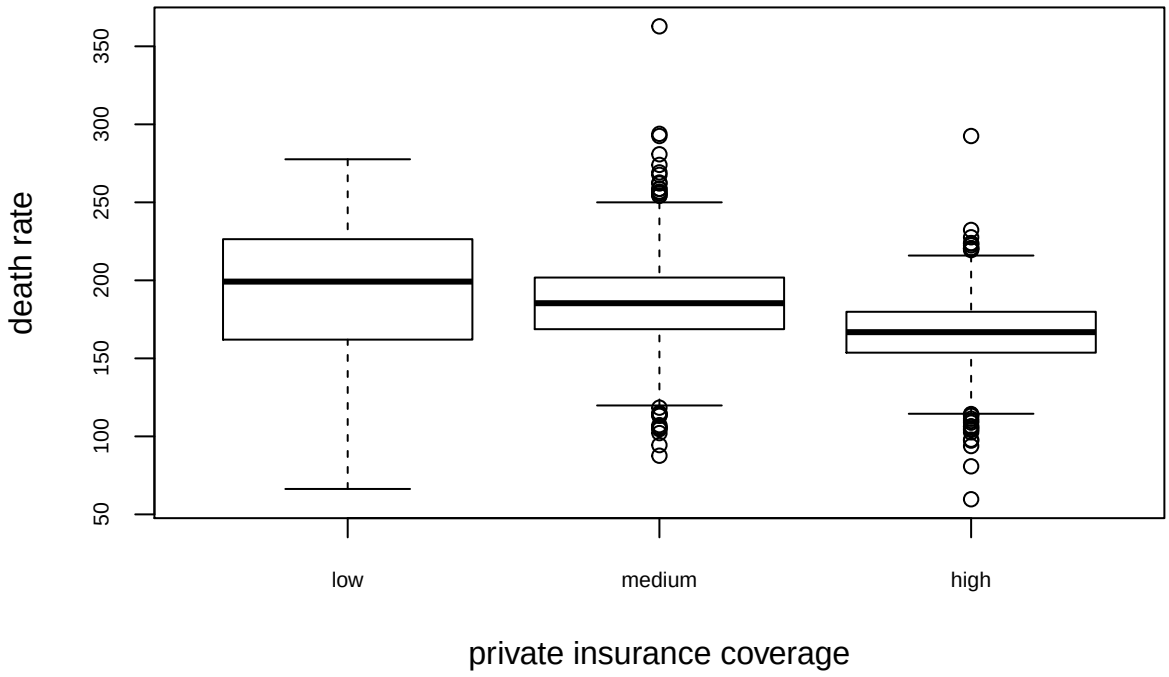
```
## [1] -0.3860655
```

Both from the plot and from the correlation value (-0.39) we can see that they're in agreement with our original hypothesis that mortality rates ared lower for the populations with higher percentage of private insurance coverage. The relationship does appear to be linear from about 40% of coverage onward (this is where the majority of observations seem to fall). At the lower end of the graph, the spread of values is much higher. Despite showing the overall trend, the scatterplot is quite noisy, so we might want to confirm our conclusion by generating boxplots for different categories of coverage. First, we'll split the range of PctPrivateCoverage variables into three bins and label them as "low", "medium", and "high" brackets of private insurance coverage. We then will build three separate boxplots for these categories and see how they're distributed relative to deathRate.

```
levels(cut(cancer.df$PctPrivateCoverage, 3, include.lowest=TRUE))
```

```
## [1] "[22.2,45.6]" "(45.6,69]"   "(69,92.4]"
```

```
boxplot(deathRate ~ cut(PctPrivateCoverage, 3, include.lowest=TRUE,
        labels=c("low", "medium", "high")),
        data = cancer.df,
        cex.axis = .7,
        main = "Death Rate for different levels of private insurance coverage",
        xlab = "private insurance coverage", ylab = "death rate")
```



The boxplot shows a clear downward trend from the "medium" to "high" category, with the majority of values clustered around the median. The "low" category boxplot, on the other hand, has a much wider spread of data points. We might conclude, therefore, that the effect of private insurance on mortality rates is only noticable for the percantage of coverage which is above certain threshold (~40%). The "medium" category also includes the high death rate outlier we've identified earlier (>350). Therefore, the high mortality rate can't be explained by the inadequate private insurance coverage.

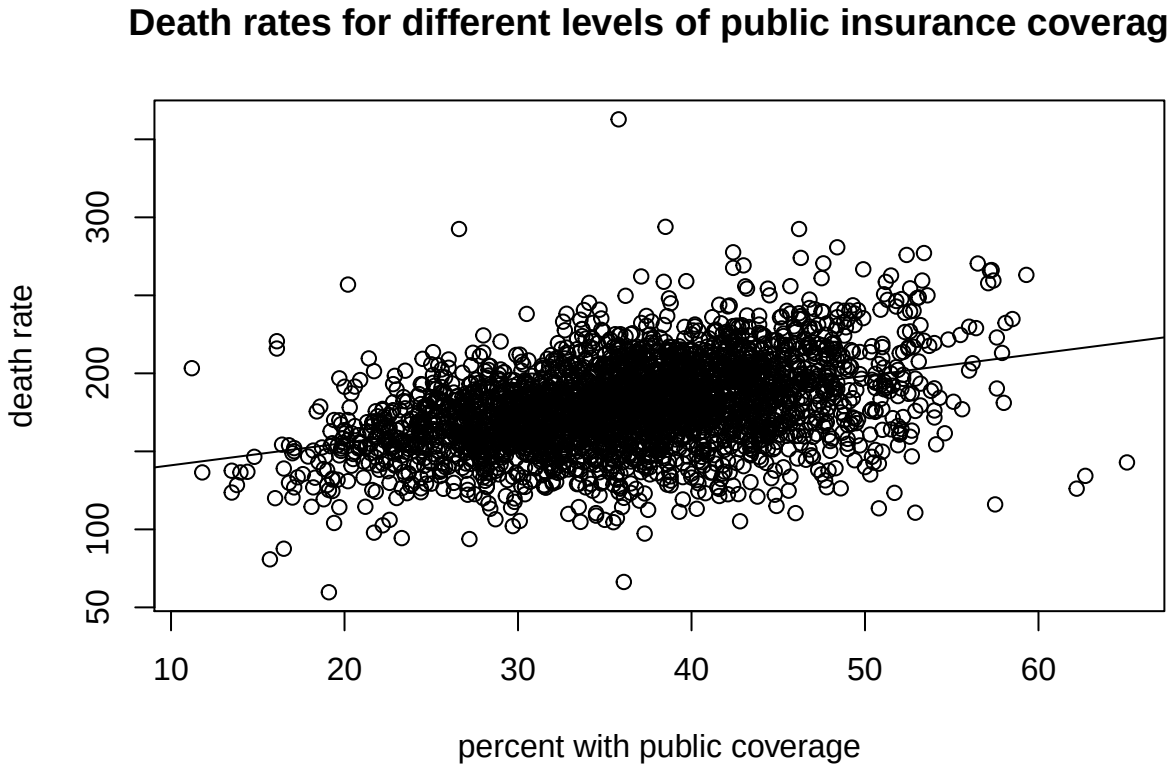
Summary of observations:

1. There's a mild negative correlation between cancer mortality rates and access to the private insurance coverage
2. The effect of negative correlation becomes noticable only after the coverage percentage reaches ~40%. Below this point, the data spread is much wider and the effect of private coverage is not obvious.

Mortality rates for different levels of public insurance coverage

We now explore whether public insurance coverage has a similar effect on cancer mortality rates. We repeat the same steps of data analysis we've performed for the private insurance variable:

```
plot(cancer.df$PctPublicCoverage, cancer.df$deathRate,
     xlab = "percent with public coverage", ylab = "death rate",
     main = "Death rates for different levels of public insurance coverage")
abline(lm(cancer.df$deathRate ~ cancer.df$PctPublicCoverage))
```



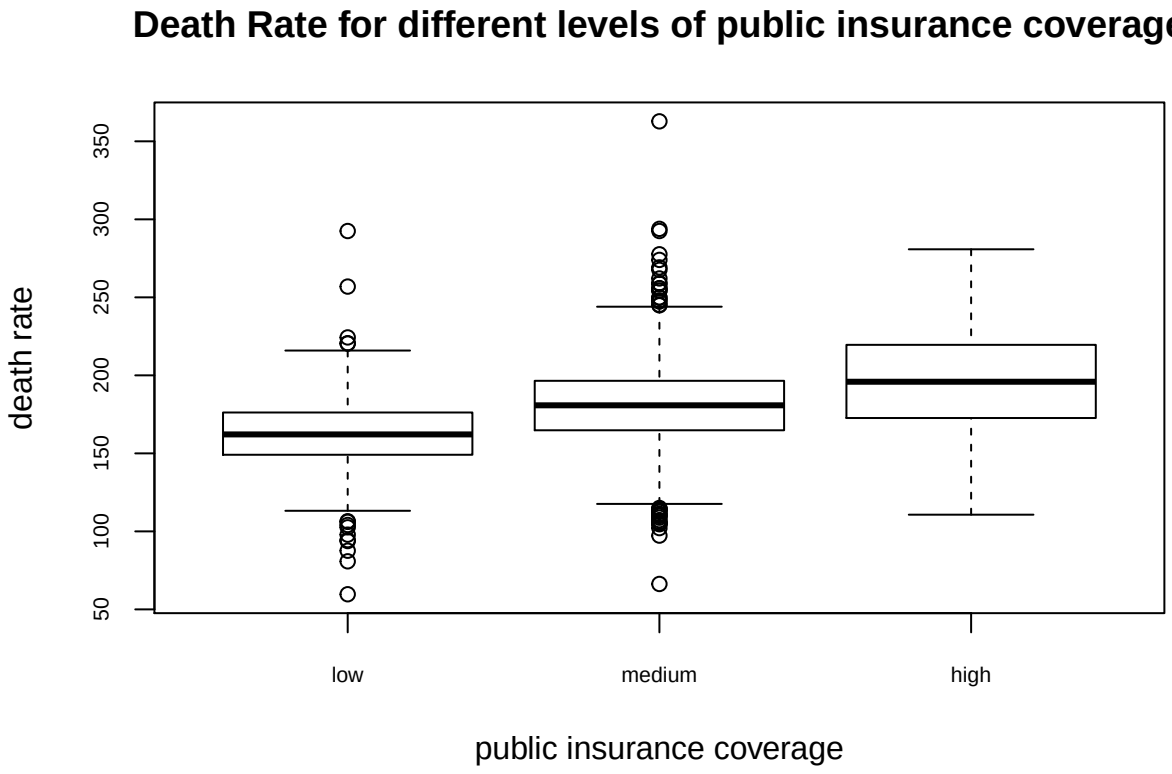

```r
cor(cancer.df$deathRate, cancer.df$PctPublicCoverage)
```

```
## [1] 0.4045717
```

```r
levels(cut(cancer.df$PctPublicCoverage, 3, include.lowest=TRUE))
```

```
## [1] "[11.1,29.2]" "(29.2,47.1]" "(47.1,65.2]"
```

```r
boxplot(deathRate ~ cut(PctPublicCoverage, 3, include.lowest=TRUE,
        labels=c("low", "medium", "high")),
        data = cancer.df,
        cex.axis = .7,
        main = "Death Rate for different levels of public insurance coverage",
        xlab = "public insurance coverage", ylab = "death rate")
```



Contrary to our expectations, we see the directly opposite relationship between public insurance coverage and cancer mortality rates. The values are positively correlated and the correlation's absolute value is even higher than the one we calculated for private insurance coverage. There's also no salient threshold effect we observed earlier: the relationship appears to be linear throughout the entire range of coverage percentage.

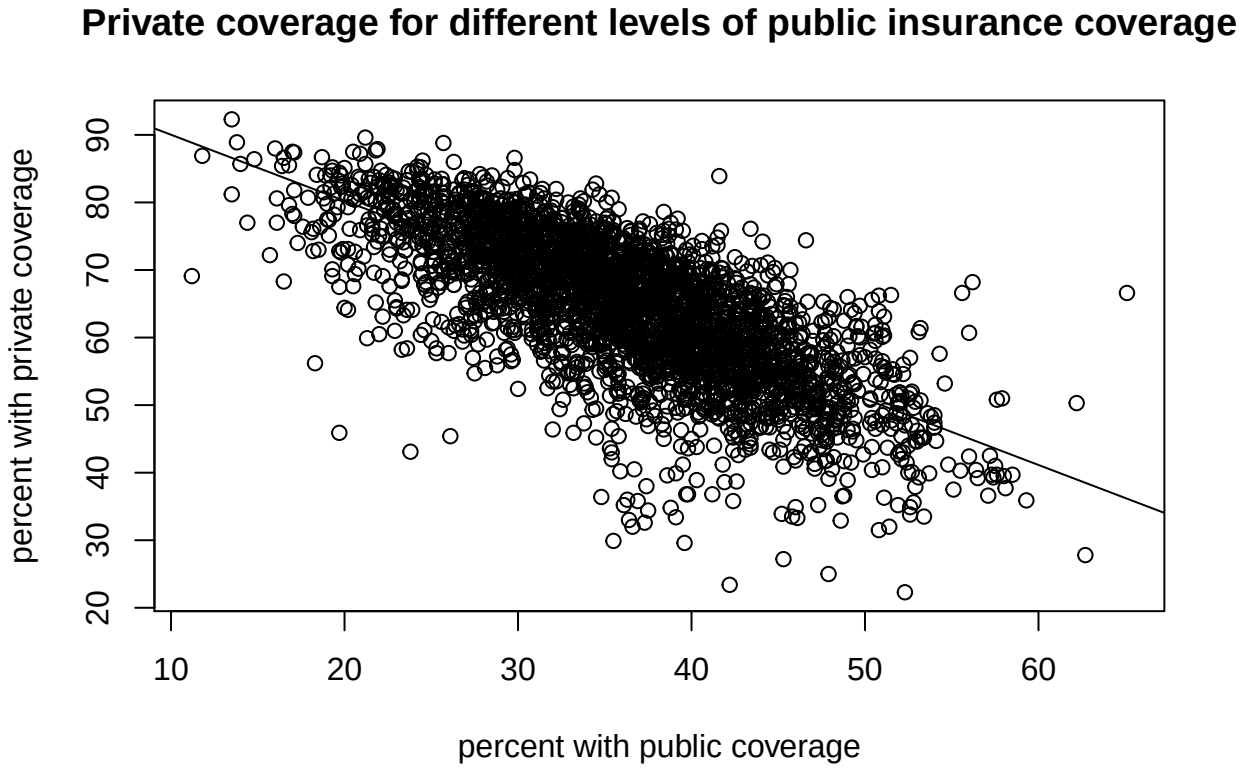
Summary of observations:

1. There's a noticeble positive correlation between cancer mortality rates and availability of public insurance coverage
2. The relationship is very close to the linear one throughout the entire range of coverage's percentages

Relationship between private and public insurance coverage

We will now explore if there is any meaningful relationship between private and public insurance coverage. As in the earlier steps of our investigation, we generate a scatterplot and box plots for these variables, and compute the correlation value:

```
plot(cancer.df$PctPublicCoverage, cancer.df$PctPrivateCoverage,
     xlab = "percent with public coverage", ylab = "percent with private coverage",
     main = "Private coverage for different levels of public insurance coverage")
abline(lm(cancer.df$PctPrivateCoverage ~ cancer.df$PctPublicCoverage))
```



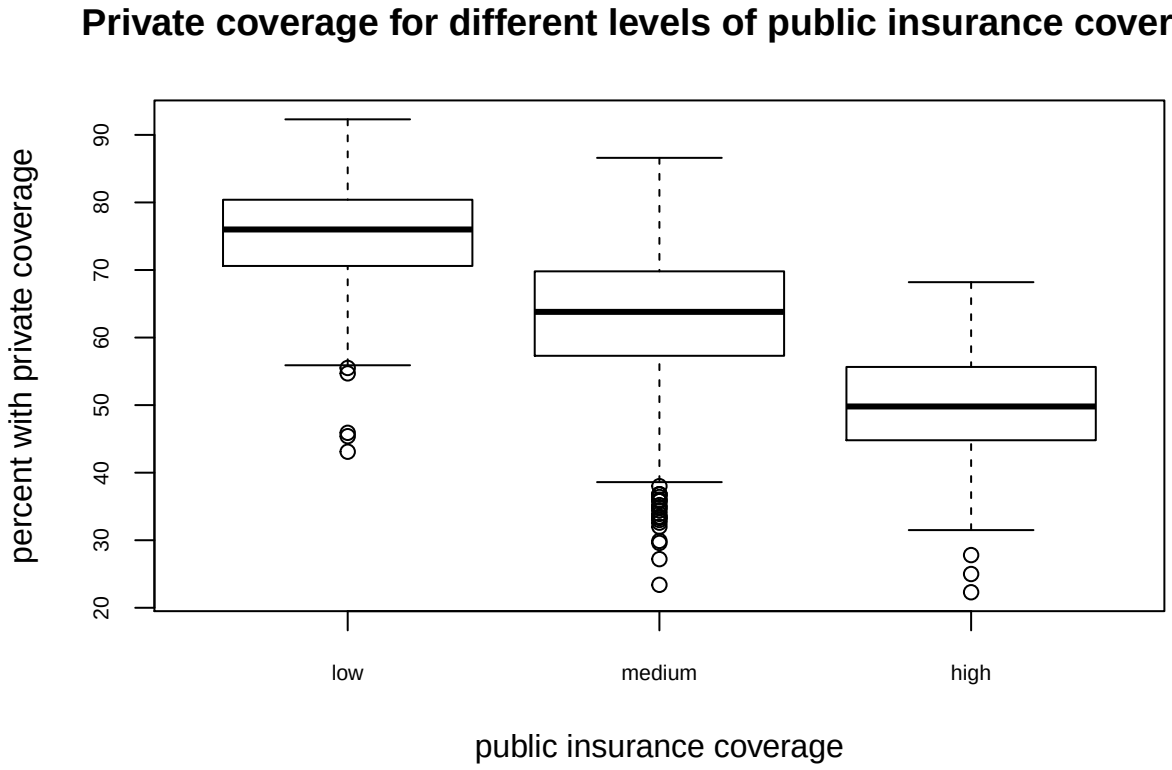

```
cor(cancer.df$PctPrivateCoverage, cancer.df$PctPublicCoverage)
```

```
## [1] -0.7200115
```

```
levels(cut(cancer.df$PctPublicCoverage, 3, include.lowest=TRUE))
```

```
## [1] "[11.1,29.2]" "(29.2,47.1]" "(47.1,65.2]"
```

```
boxplot(PctPrivateCoverage ~ cut(PctPublicCoverage, 3, include.lowest=TRUE,
        labels=c("low", "medium", "high")),
        data = cancer.df,
        cex.axis = .7,
        main = "Private coverage for different levels of public insurance coverage",
        xlab = "public insurance coverage", ylab = "percent with private coverage")
```



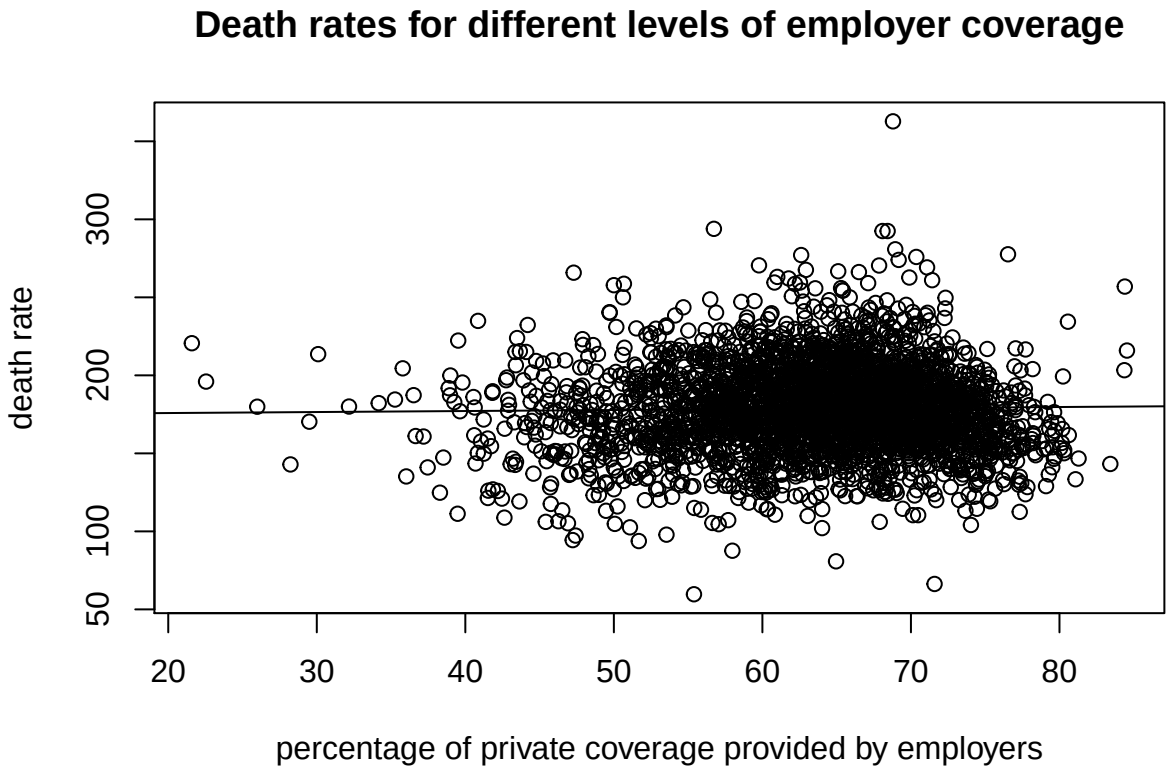
Summary of observations:

1. There's a strong negative correlation between private and public insurance coverage
2. The majority of observations cluster around ordinary least squares regression line, emphasizing linear relationship between the two variables

Mortality rates for different levels of employer-sponsored private coverage

Finally, let's see if the relative portion of employer-sponsored private insurance coverage has any relationship with cancer mortality rates.

```
plot(cancer.df$EmpSponsoredPct, cancer.df$deathRate,
     xlab = "percentage of private coverage provided by employers",
     ylab = "death rate",
     main = "Death rates for different levels of employer coverage")
abline(lm(cancer.df$deathRate ~ cancer.df$EmpSponsoredPct))
```




```
cor(cancer.df$deathRate, cancer.df$EmpSponsoredPct)
```

```
## [1] 0.01885173
```

Summary of observations:

1. From the data analysis above, we don't detect any noticable relationships between the cancer mortality rates and the composition of the private insurance coverage.

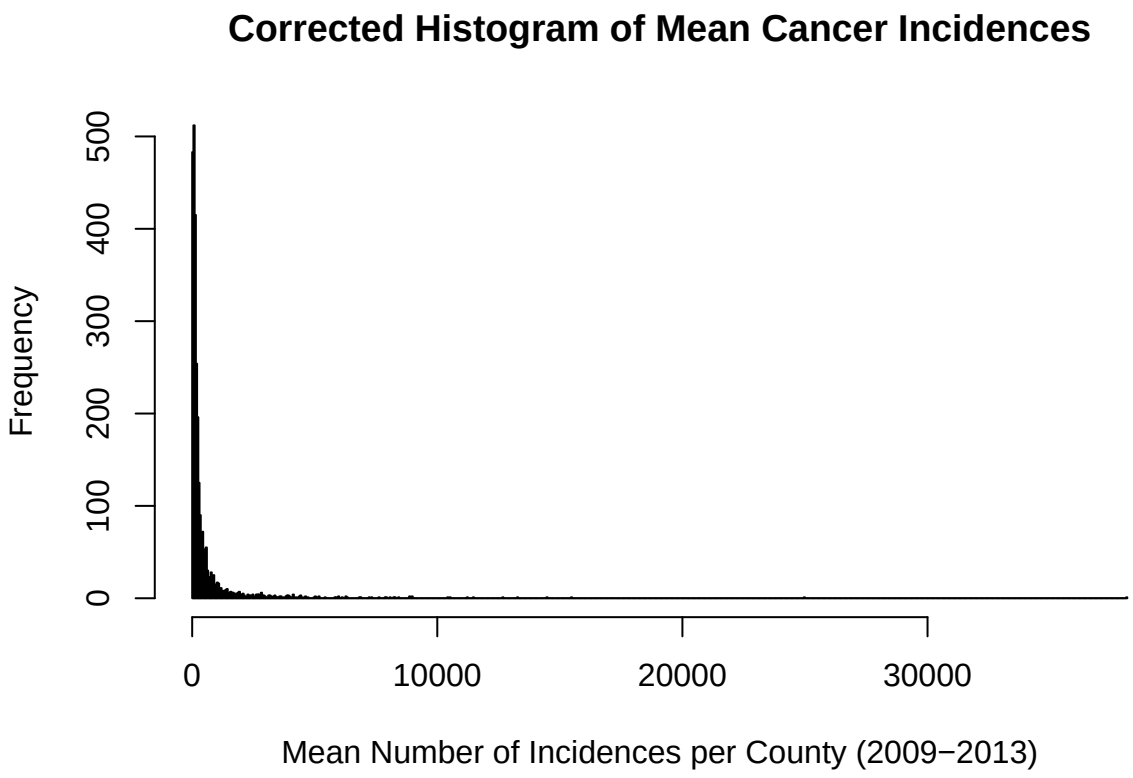
Analysis of Secondary Effects

Since we have seen that private and public insurance have opposite relationships with cancer mortality rates and highly negatively correlated between each other, we now must explore other variables that could be influencing these relationships.

We will start with a univariate analysis of other selected variables

Analysis of avgAnnCount

```
hist(cancer.df$avgAnnCount, breaks="fd", xlab = "Mean Number of Incidences per County (2009-2013)", ylab
```



That extremely right-skewed distribution is an indicative that we could use a log() transformation in this variable. Let's see how the variable is before the transformation.

```
summary(cancer.df$avgAnnCount)
```

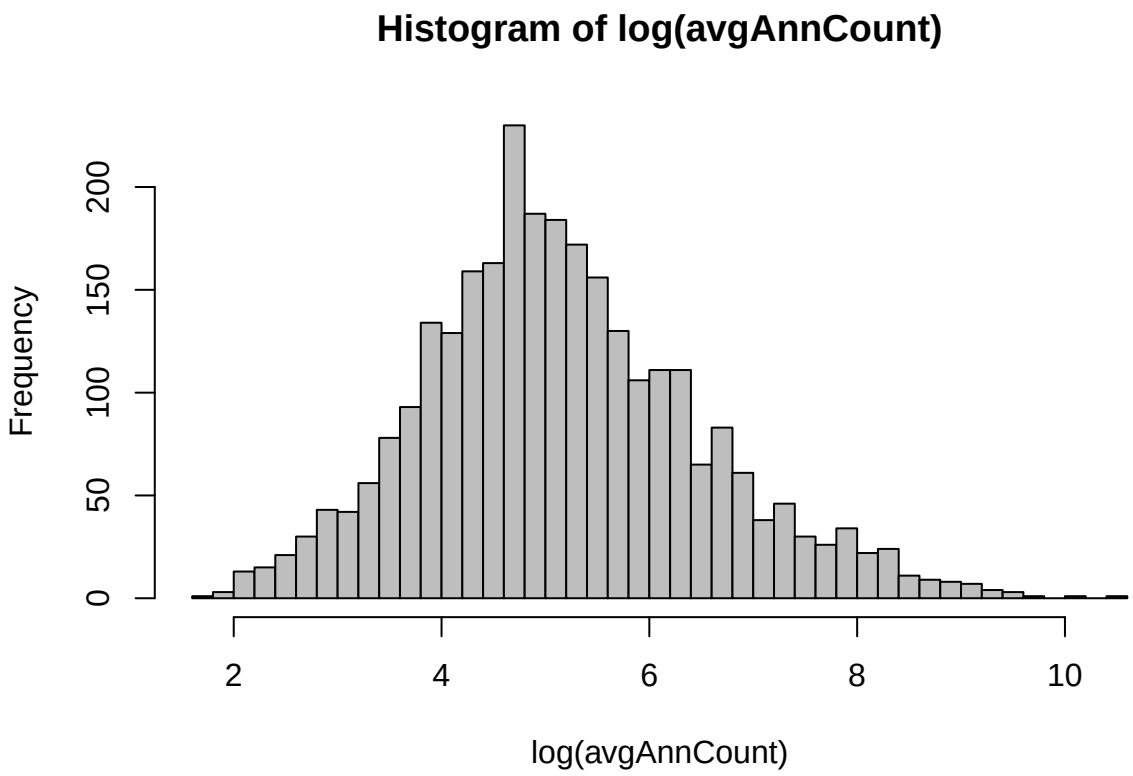
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       6      71     153     508     396   38150     206
```

Now to the analysis of the log(avgAnnCount)

```
summary(log(cancer.df$avgAnnCount))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.792   4.263   5.030   5.158   5.981  10.550     206
```

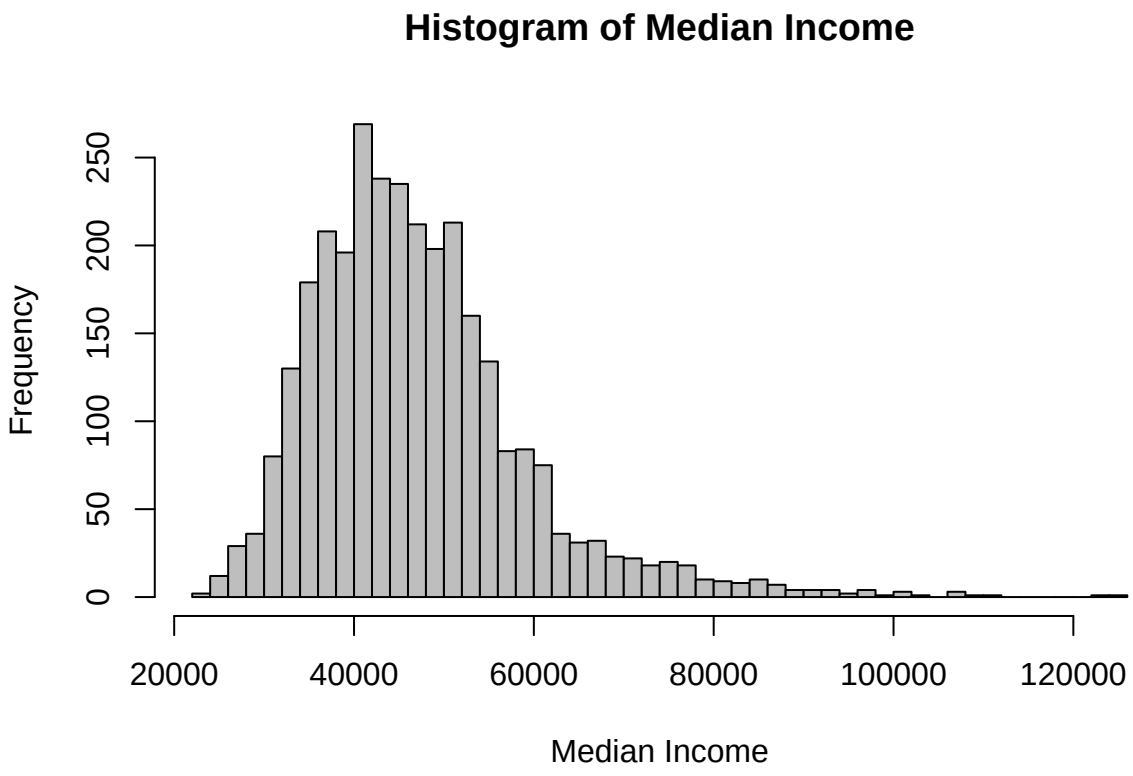
```
hist(log(cancer.df$avgAnnCount), breaks="fd", xlab = "log(avgAnnCount)", ylab="Frequency", main = "Hist
```



The distribution of log(avgAnnCount) is fairly close to normal.

Analysis of medIncome

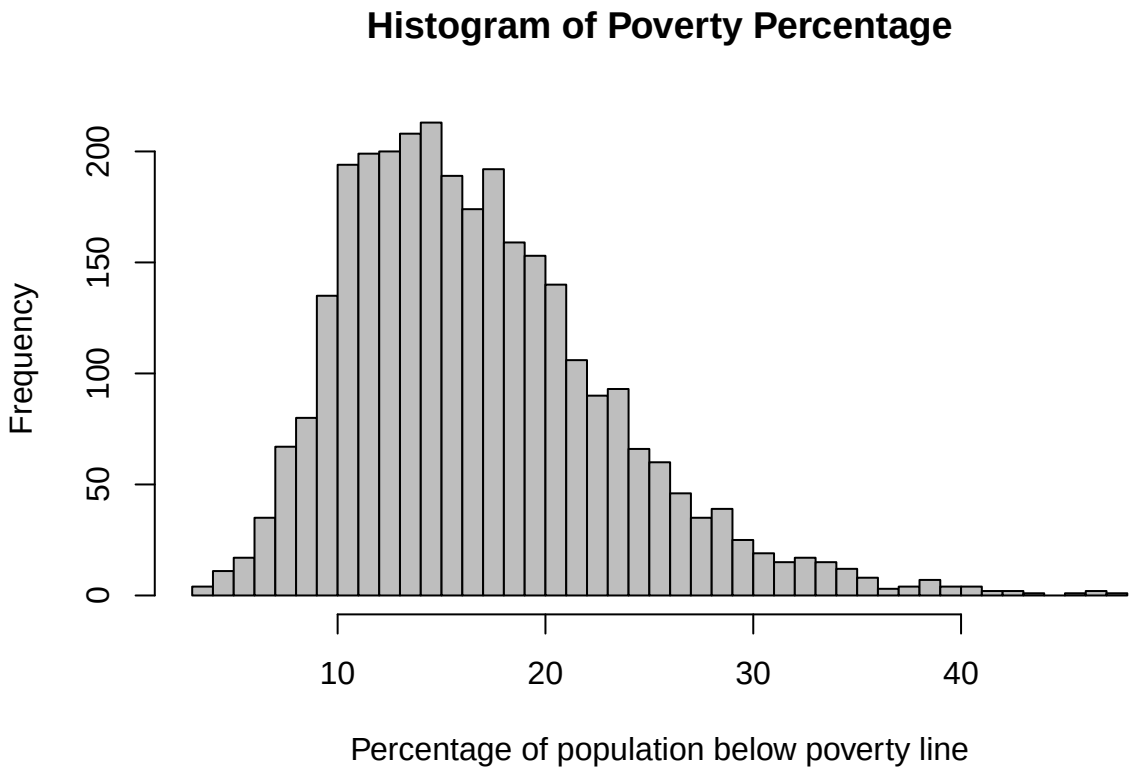
```
hist(cancer.df$medIncome, breaks="fd", xlab = "Median Income", ylab="Frequency", main = "Histogram of M
```



The distribution of Median Income is fairly normal with a slight right skew.

Analysis of povertyPercent

```
hist(cancer.df$povertyPercent, breaks="fd", xlab = "Percentage of population below poverty line", ylab=
```



The distribution of Poverty Percent is fairly normal with a slight right skew.

Bivariate Analysis of Secondary Variables

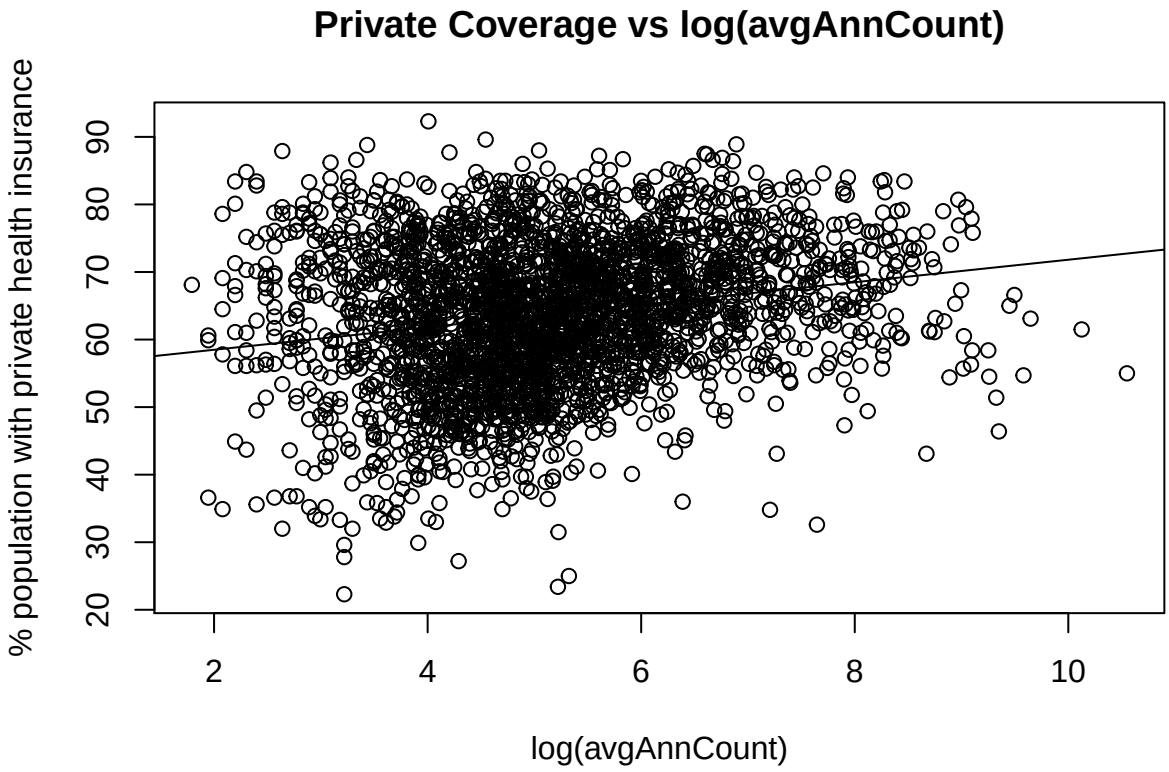
After taking a first look into other variables, and performing necessary corrections, we must now understand how they relate to the primary key variables, in order to comprehend what else might be driving the relationship previously found between different types of health insurance coverage and cancer death rates.

For each secondary variable introduced, we will analyze their relationship with the primary variables and with the output variable itself.

log(avgAnnCount)

Private Insurance Coverage

```
plot(log(cancer.df$avgAnnCount),cancer.df$PctPrivateCoverage, ylab = "% population with private health
abline(lm(cancer.df$PctPrivateCoverage[!is.na(cancer.df$avgAnnCount)] ~ log(cancer.df$avgAnnCount[!is.na
```



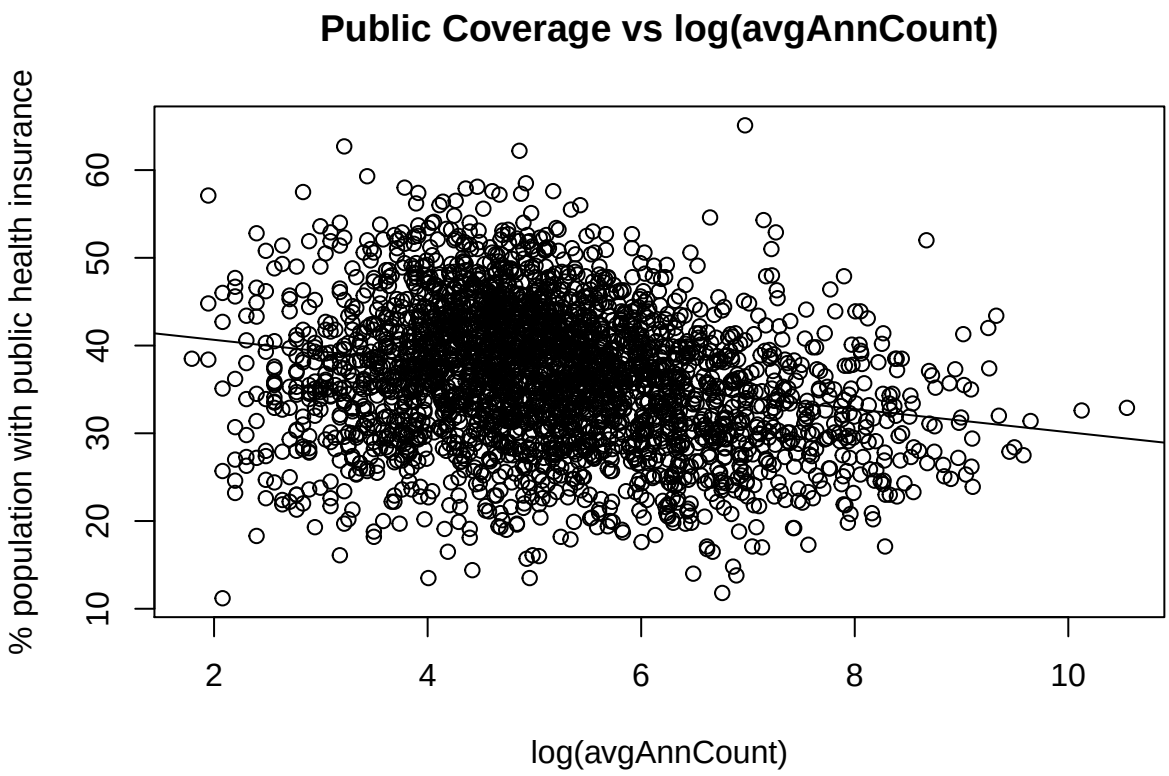

```
cor(log(cancer.df$avgAnnCount),cancer.df$PctPrivateCoverage, use = "complete.obs")
```

```
## [1] 0.2103135
```

There is a small positive correlation between the log(avgAnnCount) and the percentage of population with private health insurance coverage.

Public Insurance Coverage

```
plot(log(cancer.df$avgAnnCount),cancer.df$PctPublicCoverage, ylab = "% population with public health ins
abline(lm(cancer.df$PctPublicCoverage[!is.na(cancer.df$avgAnnCount)] ~ log(cancer.df$avgAnnCount[!is.na
```



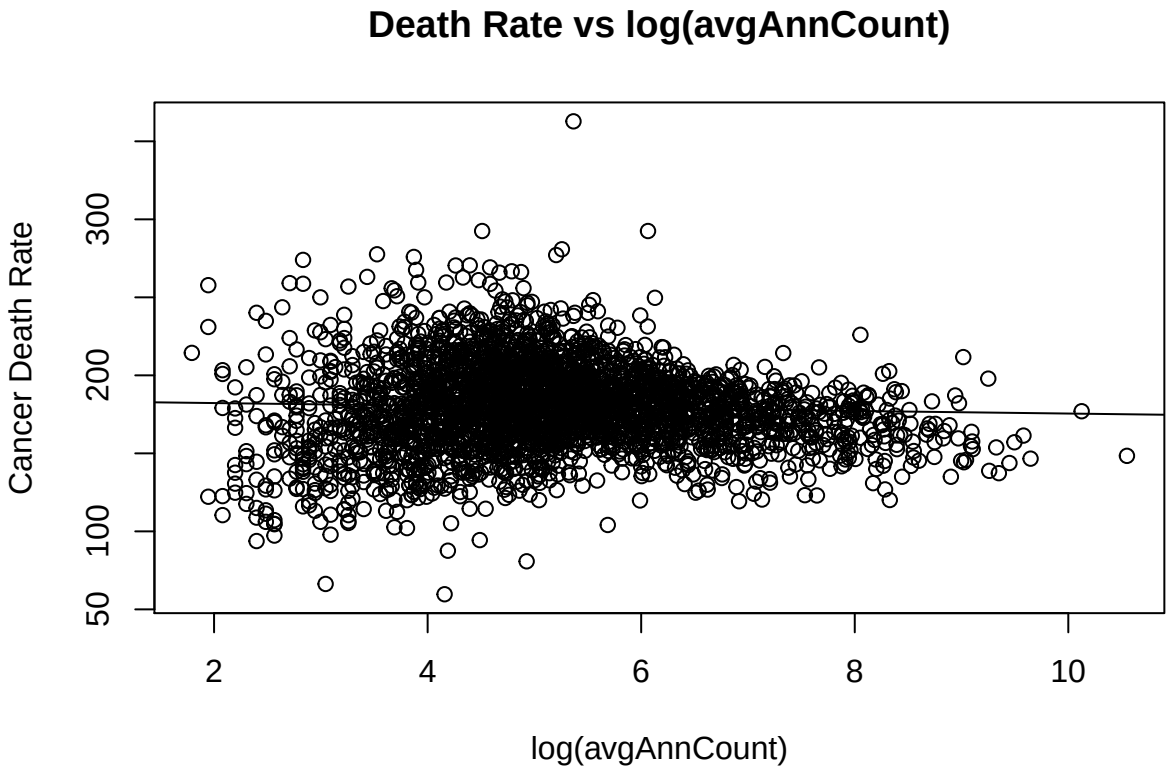

```
cor(log(cancer.df$avgAnnCount),cancer.df$PctPublicCoverage, use = "complete.obs")
```

```
## [1] -0.2241446
```

There is a small negative correlation between the log(avgAnnCount) and the percentage of population with public health insurance coverage, in a very close magnitude to the positive correlation encountrerd with the private health insurance coverage.We have seen before that, in terms of death rate, these two types of health insurance coverage have opposite behaviors. That being said, even if it is a small correlation, the fact that it presents itself in opposite ways and in similar magnitude to public and private health insurance coverage, just like death rates, indicates that we should dig deeper to check if the log(avgAnnCount) has a stronger positive correlation with death rate.

Death Rate

```
plot(log(cancer.df$avgAnnCount),cancer.df$deathRate, ylab = "Cancer Death Rate", xlab="log(avgAnnCount)
abline(lm(cancer.df$deathRate[!is.na(cancer.df$avgAnnCount)] ~ log(cancer.df$avgAnnCount[!is.na(cancer.
```




```
cor(log(cancer.df$avgAnnCount),cancer.df$deathRate, use = "complete.obs")
```

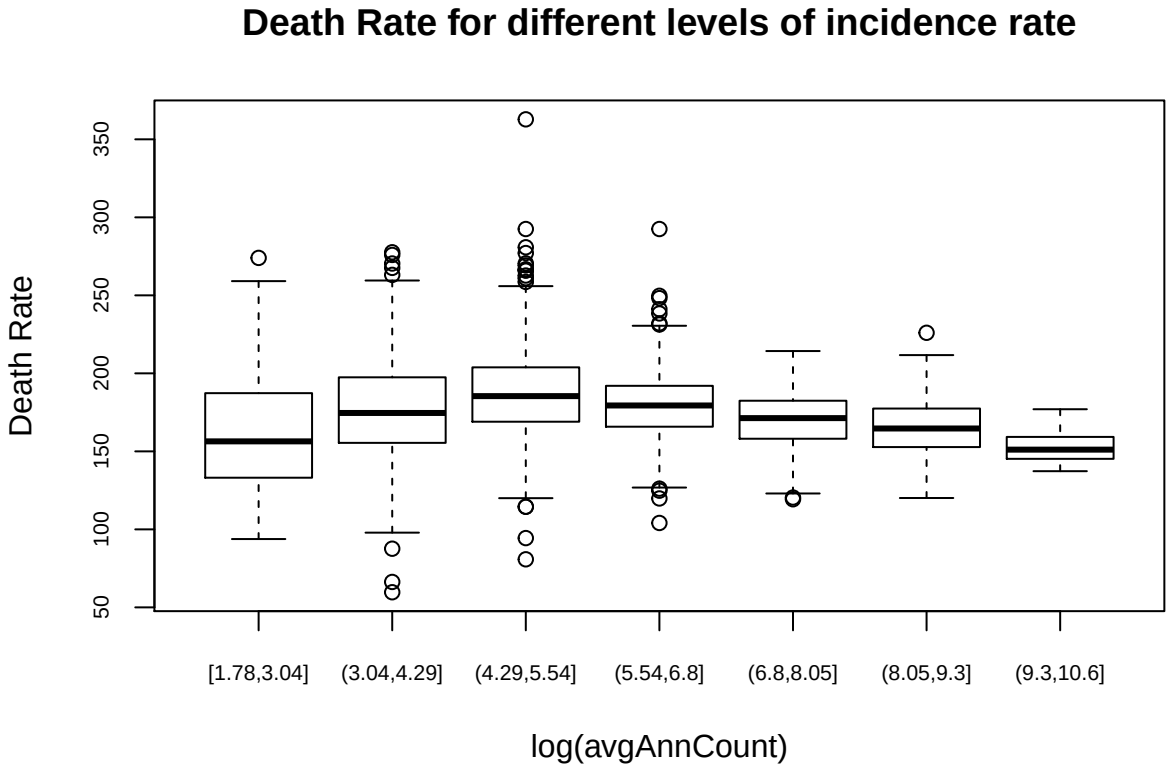
```
## [1] -0.04059046
```

At first sight, by just analyzing these charts and the correlation, it seems as we don't have much of a relation between these two variables. However, since they presented the same opposite behavior with public and private health insurance coverage, we want to take a deeper look into what might be going on.

```
levels(cut(log(cancer.df$avgAnnCount), 7, include.lowest=TRUE))
```

```
## [1] "[1.78,3.04]" "(3.04,4.29]" "(4.29,5.54]" "(5.54,6.8]"  "(6.8,8.05]"
## [6] "(8.05,9.3]"  "(9.3,10.6]"
```

```
boxplot(deathRate ~ cut(log(avgAnnCount), 7, include.lowest=TRUE),
        data = cancer.df,
        cex.axis = .7,
        main = "Death Rate for different levels of incidence rate",
        xlab = "log(avgAnnCount)", ylab = "Death Rate")
```

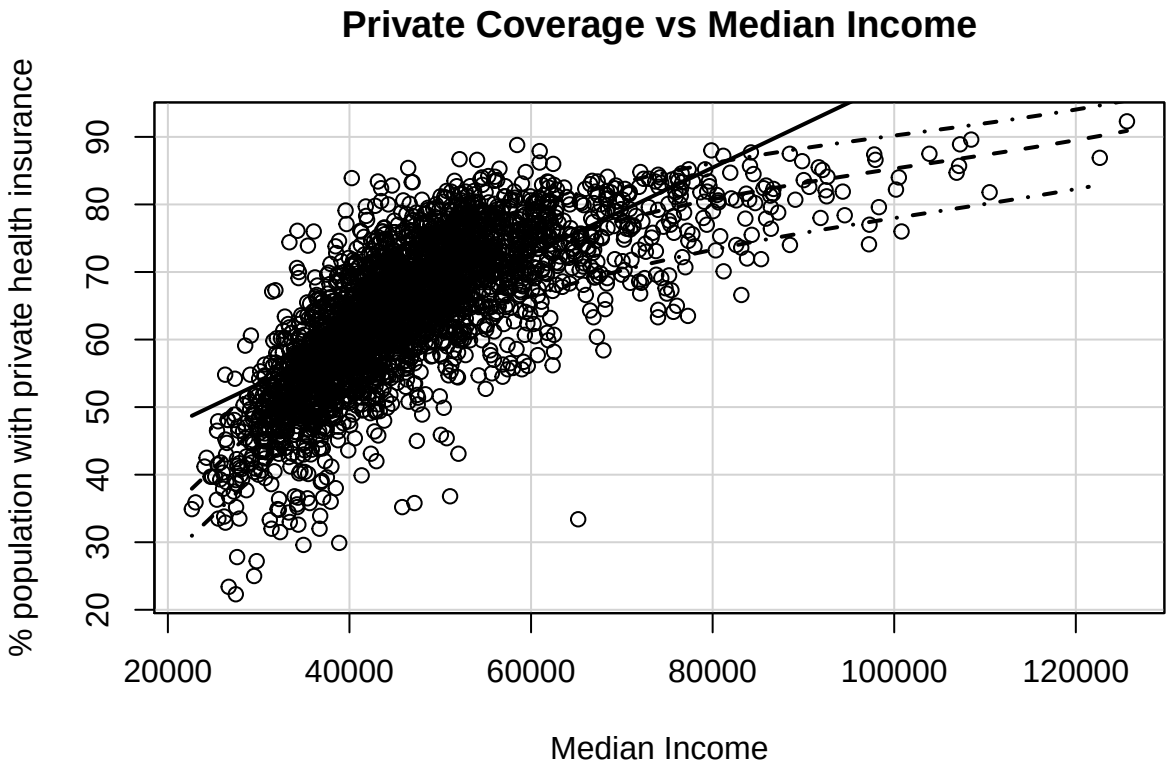


With this visualization it seems as the incidence has a positive correlation with the death rate up to a certain point. But passed that threshold (4.29, 5.54], the correlation becomes negative. One possible interpretation we had is that after a certain number of cancer reported cases, there is a more pressing need to invest in that disease, increasing the survival chances for those with it and, consequently, decreasing the death rates.

medIncome

Private Insurance Coverage

```
scatterplot(cancer.df$medIncome,cancer.df$PctPrivateCoverage, ylab = "% population with private health
```

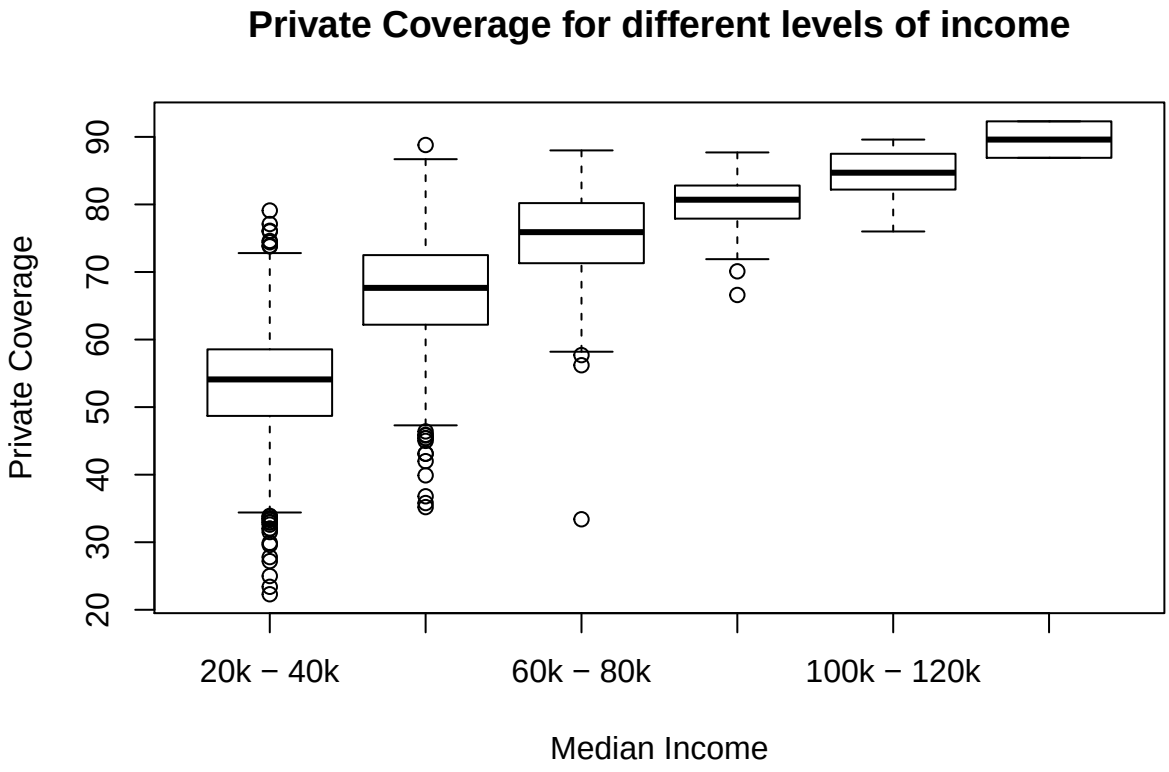



```
cor(cancer.df$medIncome,cancer.df$PctPrivateCoverage)
```

```
## [1] 0.7241748
```

By the chart presented and the strong positive correlation we atest something probably intuitevely known: populations with higher income tend to have more private health insurance coverage. We can take a deeper look by analyzing boxplots for different levels of median income:

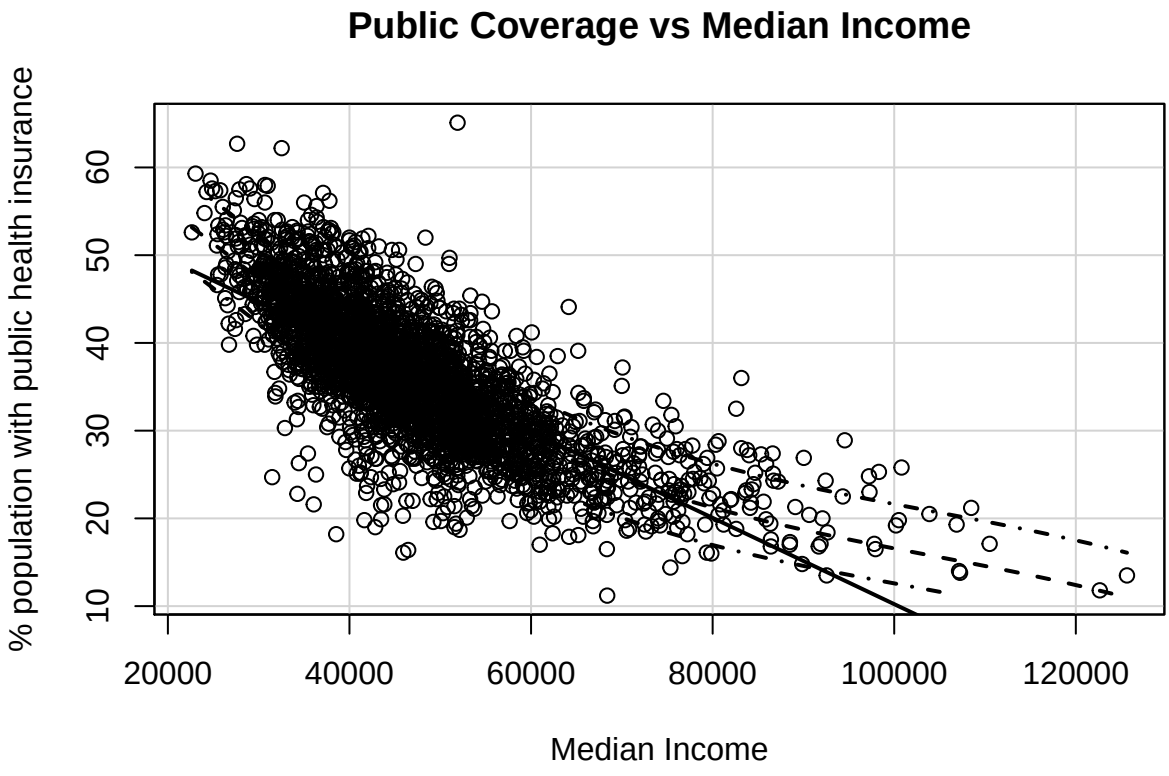
```
boxplot(cancer.df$PctPrivateCoverage ~ cut(cancer.df$medIncome, right=FALSE,seq(20000,140000,20000),labe
        main = "Private Coverage for different levels of income",
        xlab = "Median Income", ylab = "Private Coverage")
```



It confirms what we previously stated. There is a clear correlation between wealth and private health insurance coverage.

Public Insurance Coverage

```
scatterplot(cancer.df$medIncome,cancer.df$PctPublicCoverage, ylab = "% population with public health ins
```

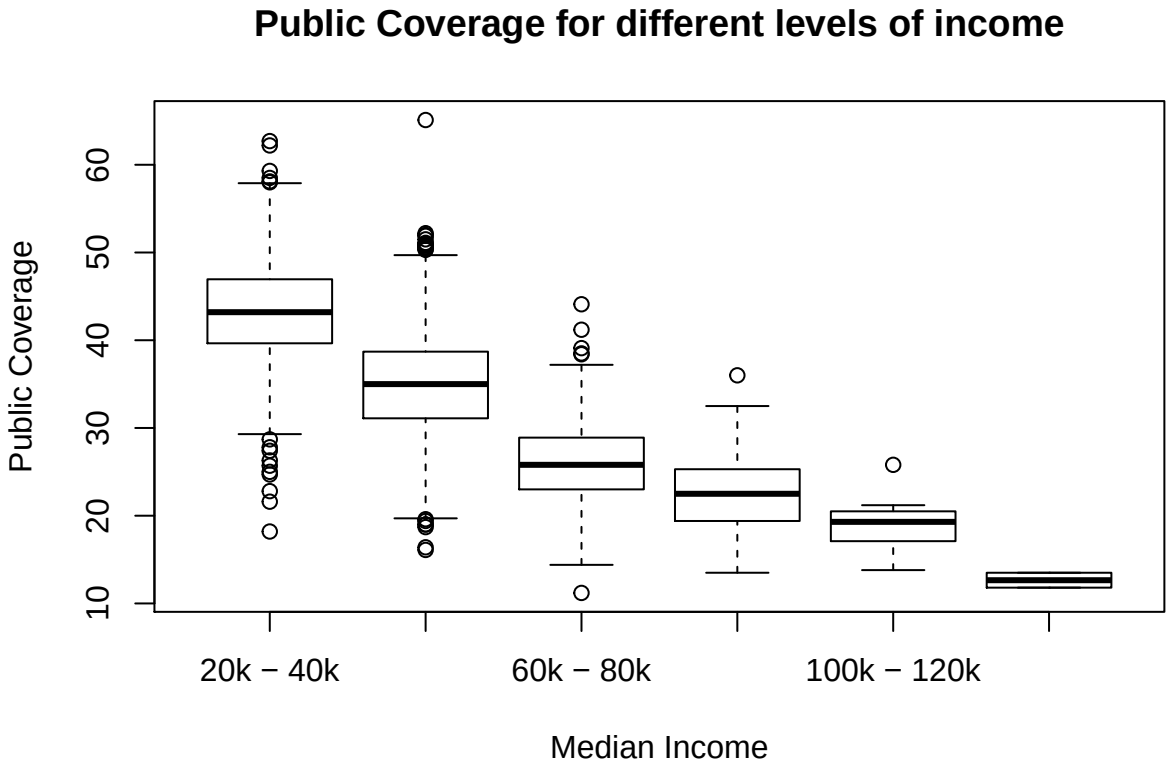


```r
cor(cancer.df$medIncome,cancer.df$PctPublicCoverage)
```

```
## [1] -0.7548218
```

Similarly to what we have seen with the Private Health Insurance Coverage, there is a clear negative correlation between Public Health Insurance Coverage and the median income. That can be interpreted that populations with lower income tend to be more dependent on Public Health Insurance, probably because the Private option is not affordable. We can take a deeper look by analyzing boxplots for different levels of median income:

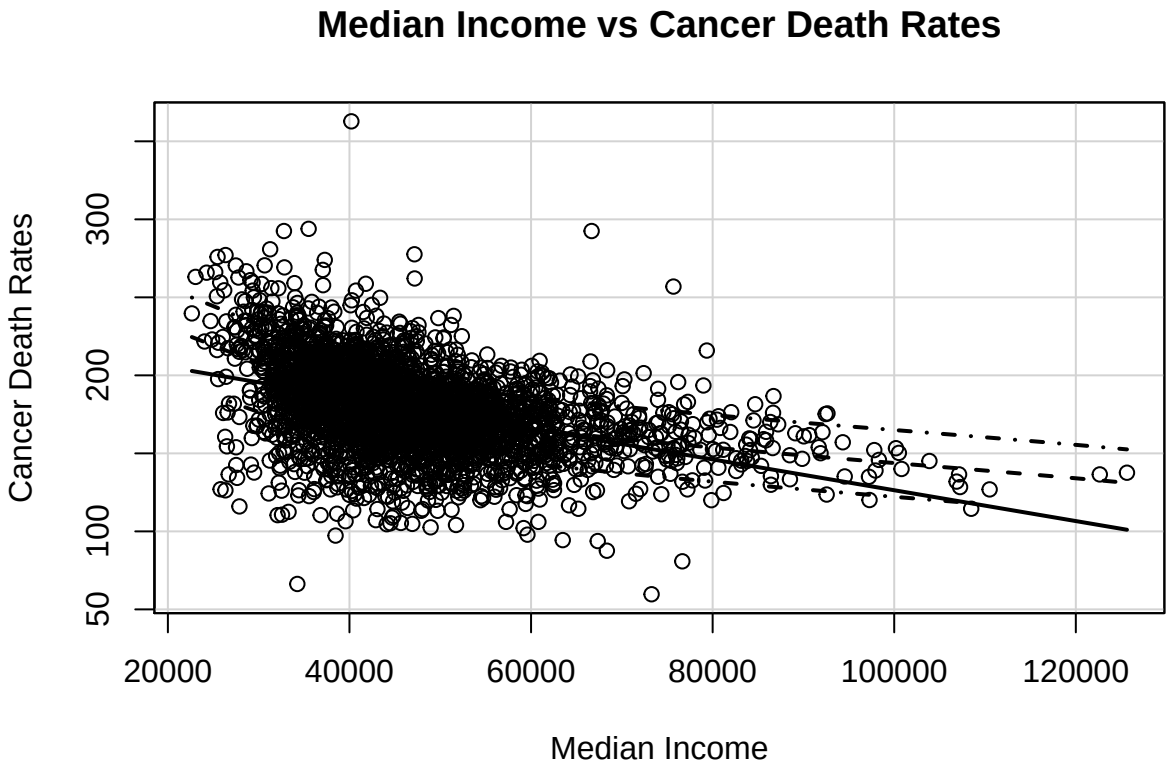
```r
boxplot(cancer.df$PctPublicCoverage ~ cut(cancer.df$medIncome, right=FALSE,seq(20000,140000,20000),labe
        main = "Public Coverage for different levels of income",
        xlab = "Median Income", ylab = "Public Coverage")
```



The boxplot only confirms what we have noticed before. For higher levels of income, the publich health insurance coverage is lower. That means this variable presents also presents the opposite behavior as death rates. The higher the median income, higher the private coverage and lower the public coverage. So we should probably check if there is a direct relation between the median income and death rates.

Death Rates

```r
scatterplot(cancer.df$medIncome,cancer.df$deathRate, ylab = "Cancer Death Rates", xlab="Median Income",
```

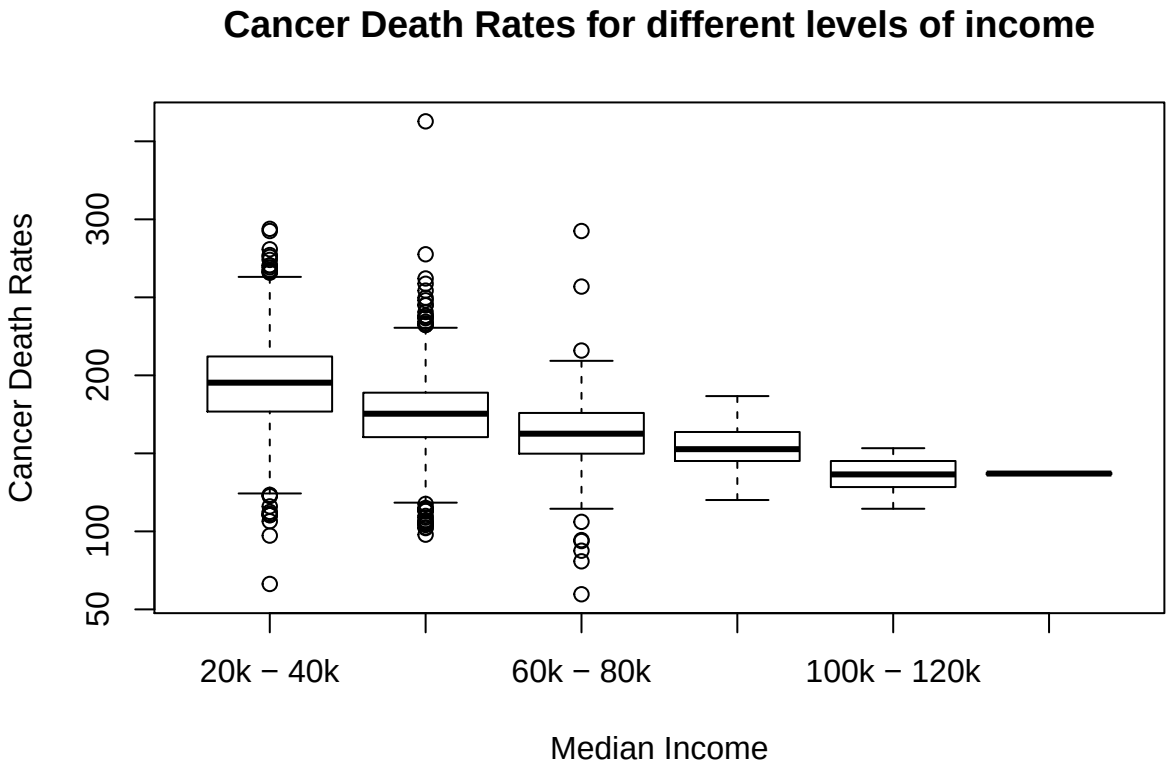



```
cor(cancer.df$medIncome,cancer.df$deathRate)
```

```
## [1] -0.4286149
```

We see that there is a stronger negative correlation between median income and death rates than private health insurance coverage and death rates, which may lead us to the hypothesis that actually socieconomic factors have more to do with death rates than the percent coverage by type of health insurance itself. Taking a deeper look by analyzing the boxplot by levels of median income may provide us with better insights

```
boxplot(cancer.df$deathRate ~ cut(cancer.df$medIncome, right=FALSE,seq(20000,140000,20000),labels = c("
        main = "Cancer Death Rates for different levels of income",
        xlab = "Median Income", ylab = "Cancer Death Rates")
```

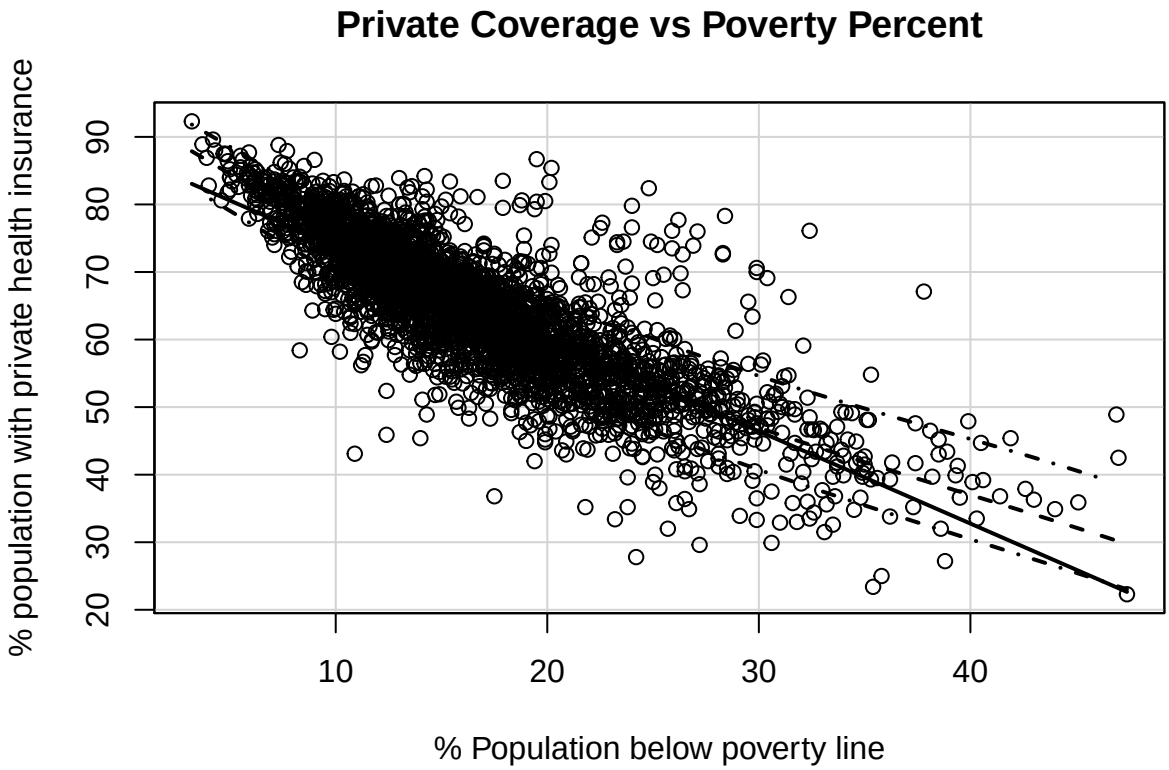


It seems to confirm our hypothesis, however, taking a deeper look into another socieconomic variable might strenghten our hypothesis.

povertyPercent

Private Insurance Coverage

```
scatterplot(cancer.df$povertyPercent,cancer.df$PctPrivateCoverage, ylab = "% population with private hea
```

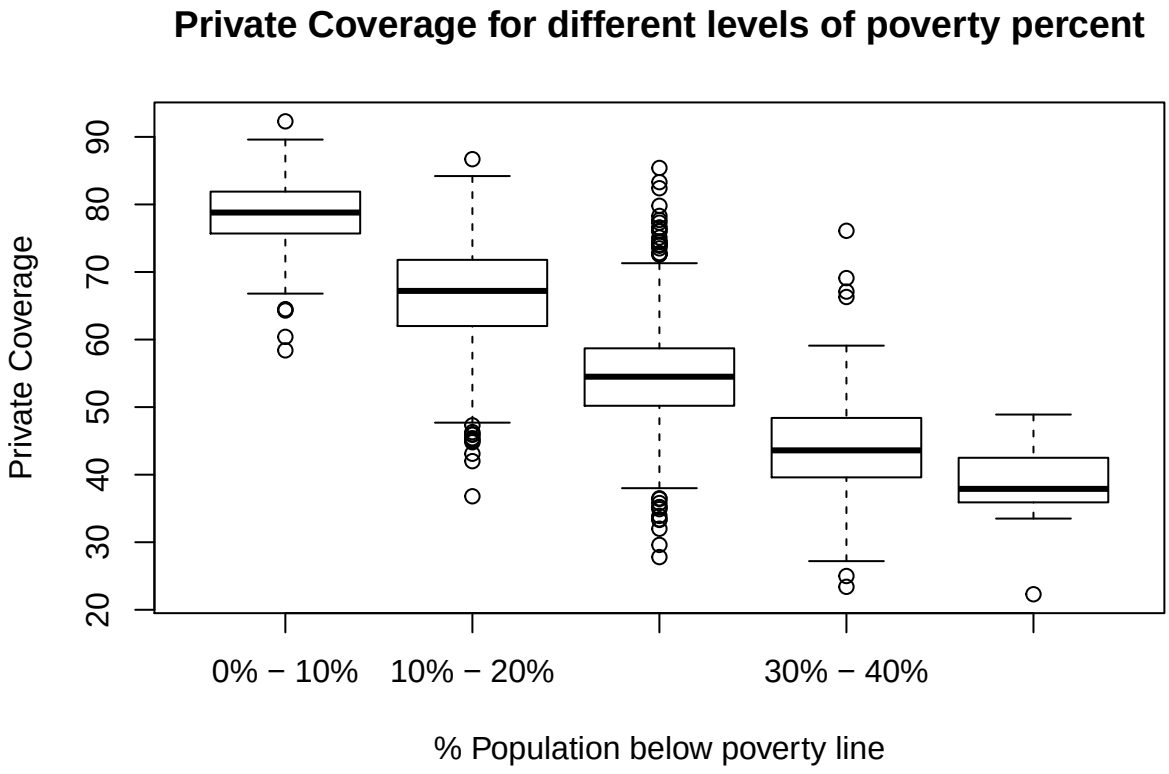



```
cor(cancer.df$povertyPercent,cancer.df$PctPrivateCoverage)
```

```
## [1] -0.8225343
```

The strongest relation we have encountered so far, we see that populations with higher percentage below poverty line tend to have less private health insurance coverage, the opposite behavior to the median income variable. Taking a look into boxplots provides us with a indicative of validity of such hypothesis:

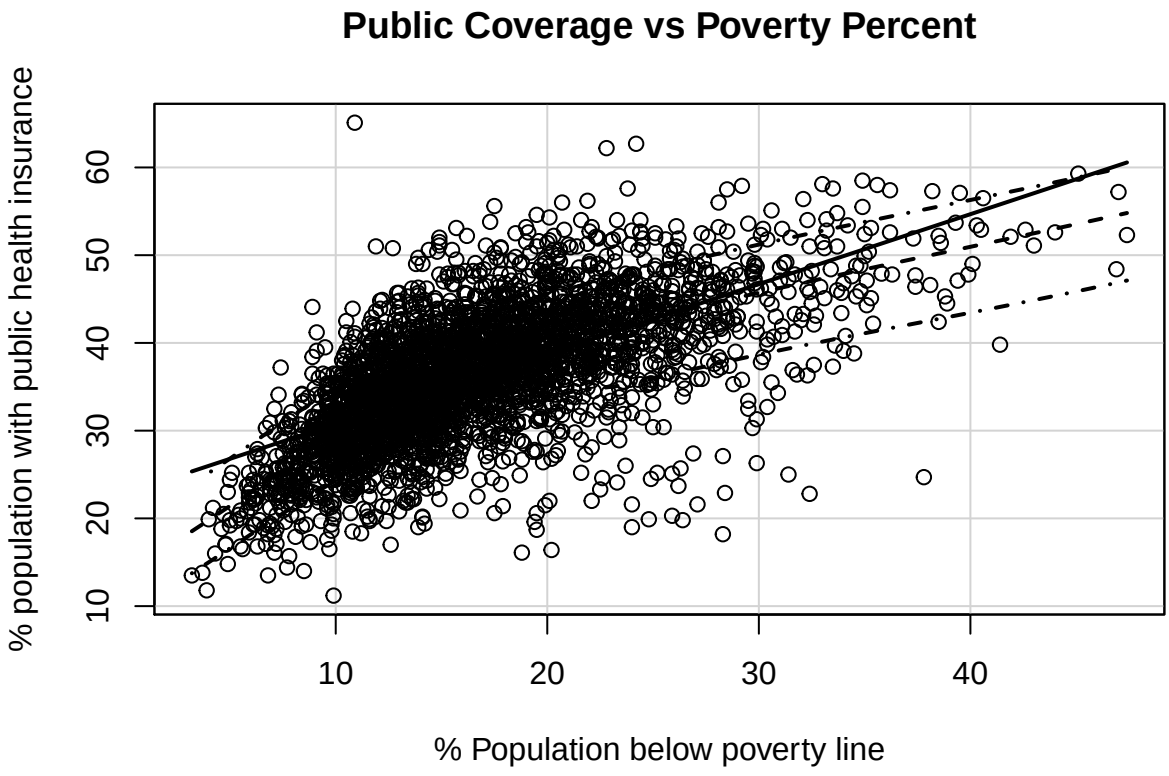
```
boxplot(cancer.df$PctPrivateCoverage ~ cut(cancer.df$povertyPercent, right=FALSE,seq(0,50,10),labels = c
        main = "Private Coverage for different levels of poverty percent",
        xlab = "% Population below poverty line", ylab = "Private Coverage")
```



As expected, it presents an opposite behavior to median income. In this case, higher the poverty (and lower the income, as we previously saw), the lower private health insurance coverage.

Public Insurance Coverage

```
scatterplot(cancer.df$povertyPercent,cancer.df$PctPublicCoverage, ylab = "% population with public heal
```

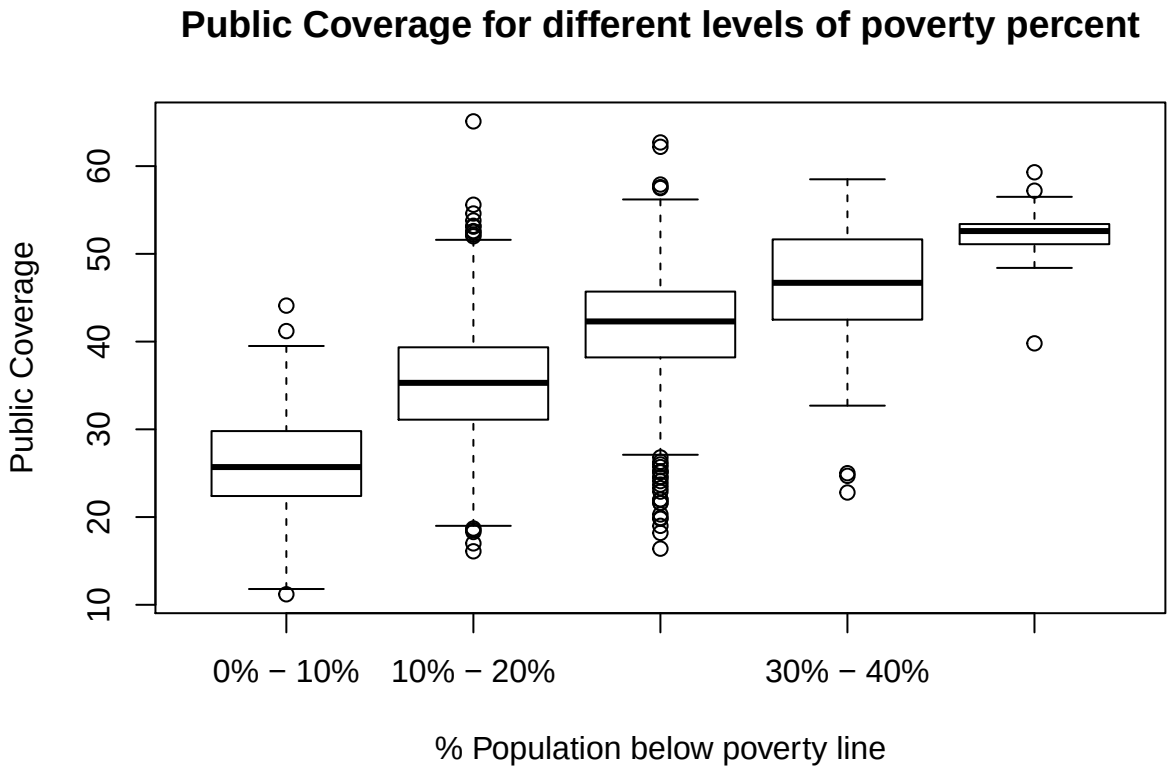


```r
cor(cancer.df$povertyPercent,cancer.df$PctPublicCoverage)
```

```
## [1] 0.6511621
```

As expected by our previous analysis, the higher the poverty, more people rely on public health insurance. A deeper look into the levels of poverty vs public health coverage might provide us with better insights:

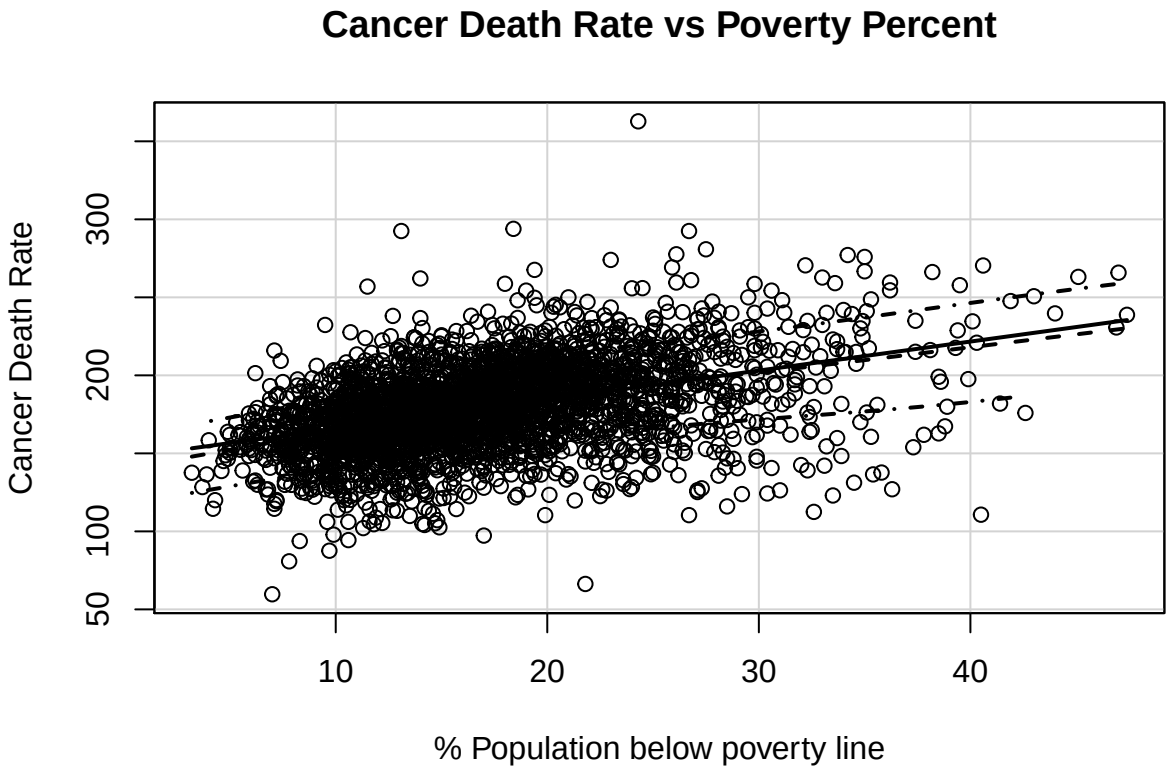
```r
boxplot(cancer.df$PctPublicCoverage ~ cut(cancer.df$povertyPercent, right=FALSE,seq(0,50,10),labels = c
        main = "Public Coverage for different levels of poverty percent",
        xlab = "% Population below poverty line", ylab = "Public Coverage")
```



While analyzing the median income, we formulated the hypothesis that maybe the death rate is driven more due to socio-economic factors than to the percentage by type of health insurance coverage. We might confirm that by analyzing also the direct relation between death rate and poverty percent, with the result strenghtening or weakening this hypothesis.

deathRate

```r
scatterplot(cancer.df$povertyPercent,cancer.df$deathRate, ylab = "Cancer Death Rate", xlab="% Populatio
```

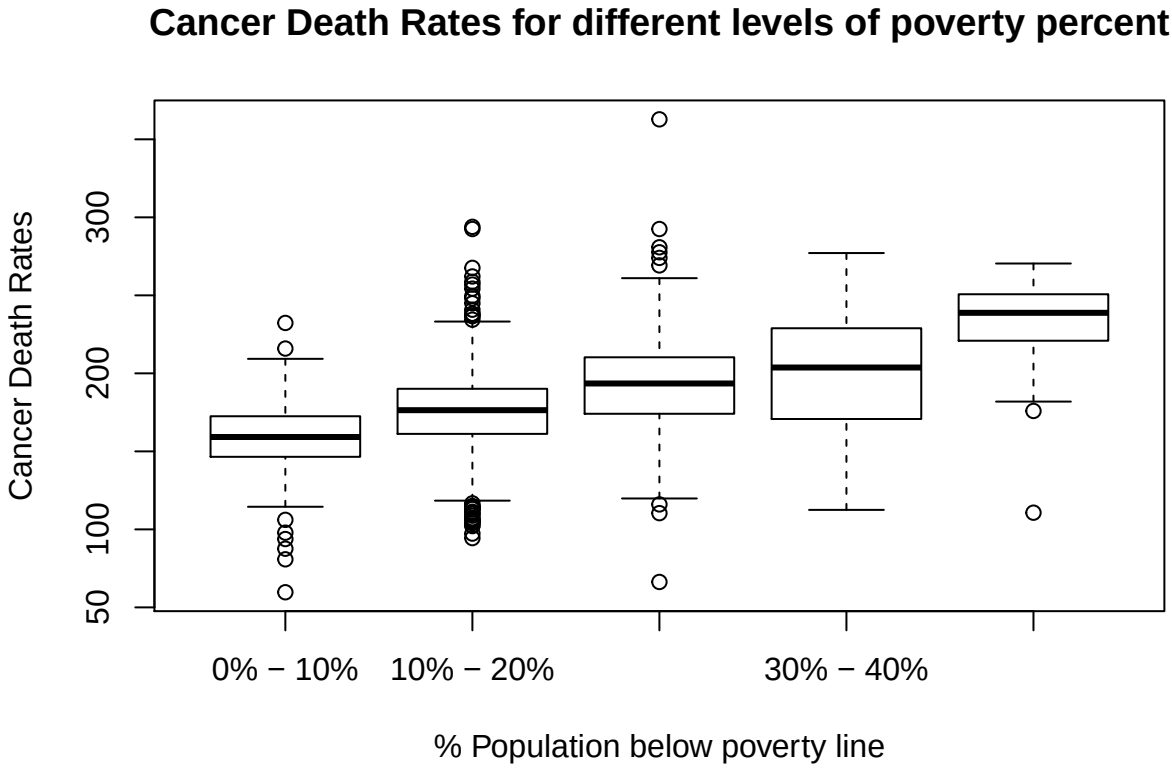



```
cor(cancer.df$povertyPercent,cancer.df$deathRate)
```

```
## [1] 0.429389
```

At first glance, we get a stronger positive relation of death rates and poverty percent than death rates and public health insurance coverage and death rates, indicating that we might be on the right track. To be more sure of it, we can make use of a boxplot by levels of poverty:

```
boxplot(cancer.df$deathRate ~ cut(cancer.df$povertyPercent, right=FALSE,seq(0,50,10),labels = c("0% - 1
        main = "Cancer Death Rates for different levels of poverty percent",
        xlab = "% Population below poverty line", ylab = "Cancer Death Rates")
```



That all seems to confirm what we have seen analyzing the median income vs the death rates. Higher income / Lower poverty counties tend to have lower death rates.

Conclusion: insurance coverage per ce doesn't improve cancer mortality rates, however, better social economic conditions seems to do so

1. The opposite behavior of public health insurance coverage and private health insurance coverage to the death rates are most likely due to an underlying factor: social economic conditions.
2. Higher income populations tend to have lower cancer death rates, and with more money, more access to private health insurance.
3. Populations with higher percentage of poverty tend to have higher cancer death rates, and poverty conditions limitate the access to private health insurance coverage, being more dependent on the public alternative.
4. Therefore, there is stronger evidence that social economic factors (income, poverty) are stronger factors in explaining the cancer death rates than health insurance per ce, being that the coverage by type of health insurance is also probably affected by these factors, explaining their opposite behaviors with death rates.