Cancer EDA

Setup

First, we load the car library, which gives us a convenient scatterplotMatrix function.

```
library(car)
```

```
## Loading required package: carData
```

```
# Load the data
cancer.df=read.csv("C:/Berkeley/W203/Lab 1/cancer.csv")
```

Data Transformation

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:

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.

Hence, let's introduce three new variables as follows:

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

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

```
nrow(cancer.df[(cancer.df$PctPublicCoverage + cancer.df$PctPrivateCoverage)>100,])
```

```
## [1] 1722
```

```
nrow(cancer.df[(cancer.df$PctPublicCoverage + cancer.df$PctPrivateCoverage)<100,])
```

```
## [1] 1313
```

```
cancer.df$PctDoubleCoverage=cancer.df$PctPublicCoverage + cancer.df$PctPrivateCoverage - 100
cancer.df$PctDoubleCoverage[cancer.df$PctDoubleCoverage < 0] = 0
summary(cancer.df$PctDoubleCoverage)
```

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

```
cancer.df$PctNoCoverage = 100 - cancer.df$PctPublicCoverage - cancer.df$PctPrivateCoverage
cancer.df$PctNoCoverage[cancer.df$PctNoCoverage < 0] = 0
summary(cancer.df$PctNoCoverage)
```

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

```
cancer.df$EmpSponsoredPct = cancer.df$PctEmpPrivCoverage / cancer.df$PctPrivateCoverage * 100
summary(cancer.df$EmpSponsoredPct)
```

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

Univariate Analysis of Key Variables

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:

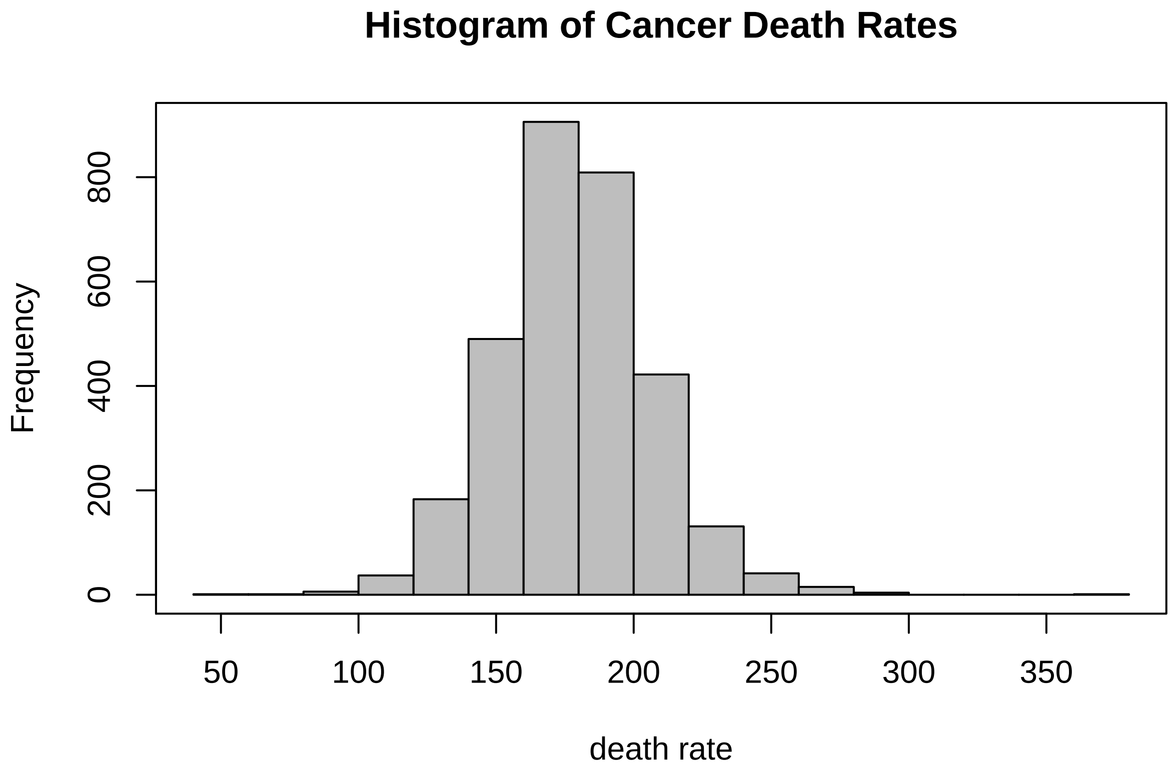
```
summary(cancer.df$deathRate)
```

```
##    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.

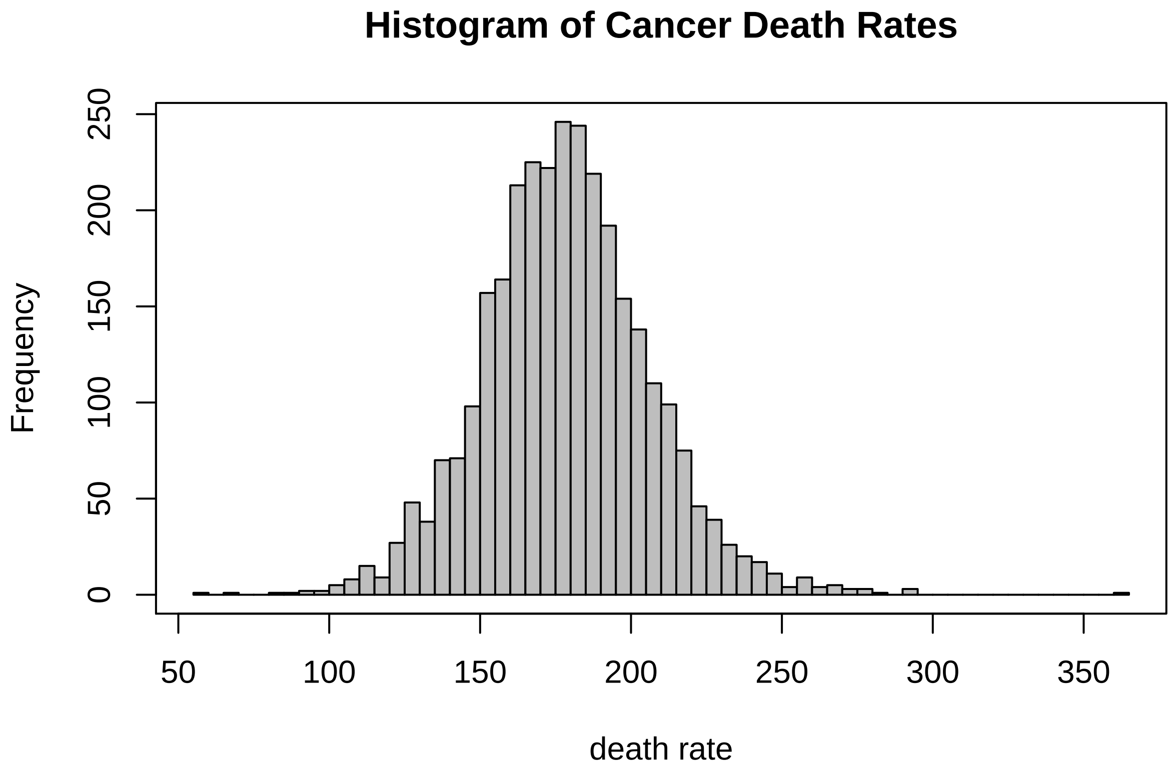
To better visualize the variable's values distribution, we plot the histogram.

```
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  33 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
##  $ PctDoubleCoverage   : num 0
##  $ PctNoCoverage       : num 4.6
##  $ EmpSponsoredPct     : num 68.8
```

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.

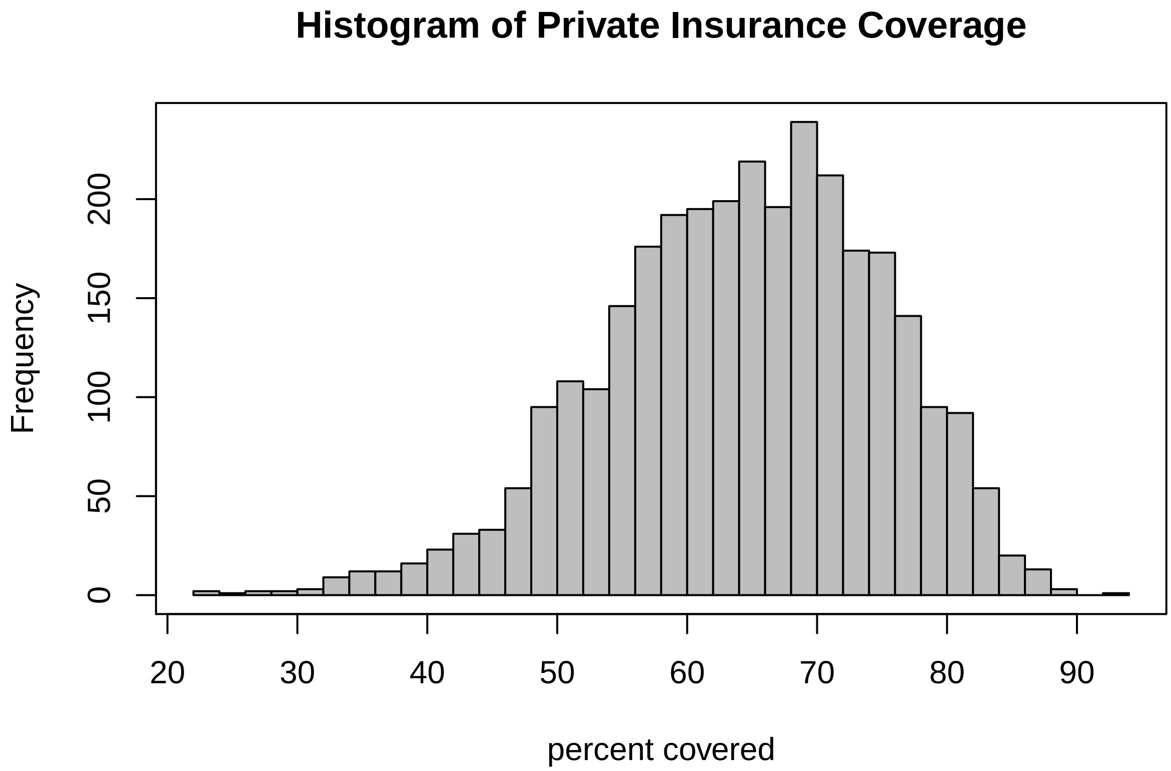
Private Insurance Coverage (PctPrivateCoverage variable)

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.

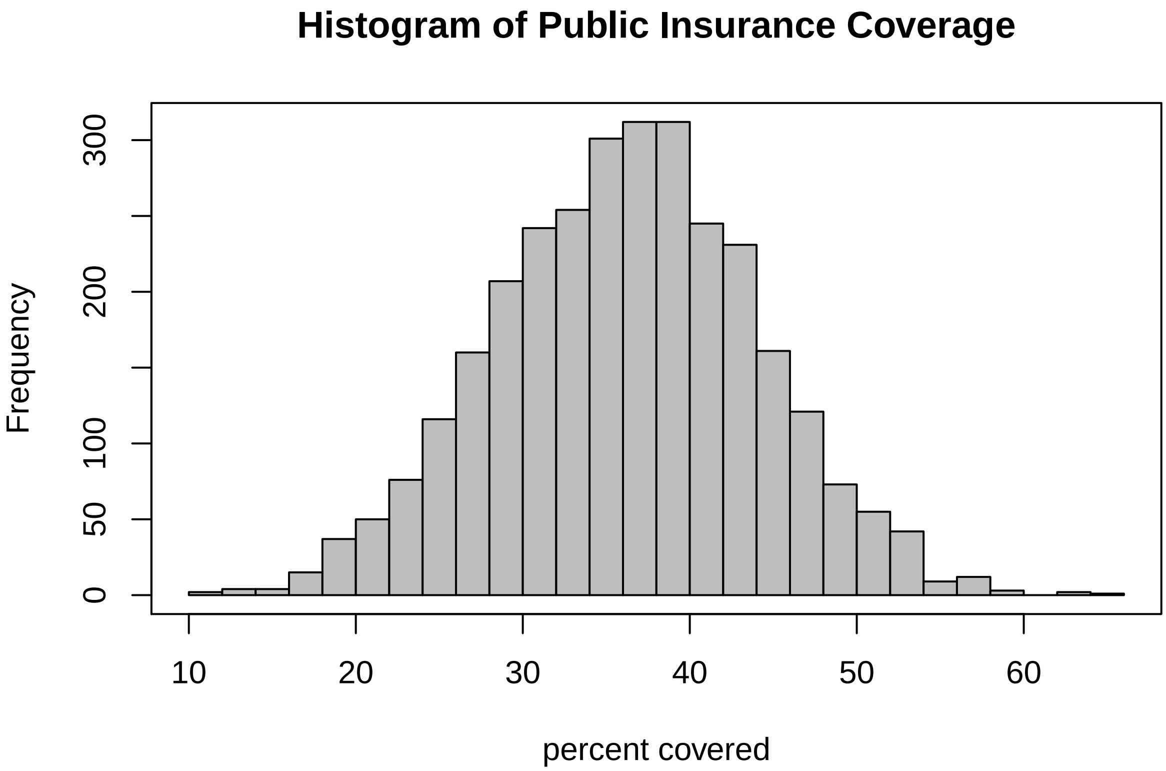
Public Insurance Coverage (PctPublicCoverage variable)

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.

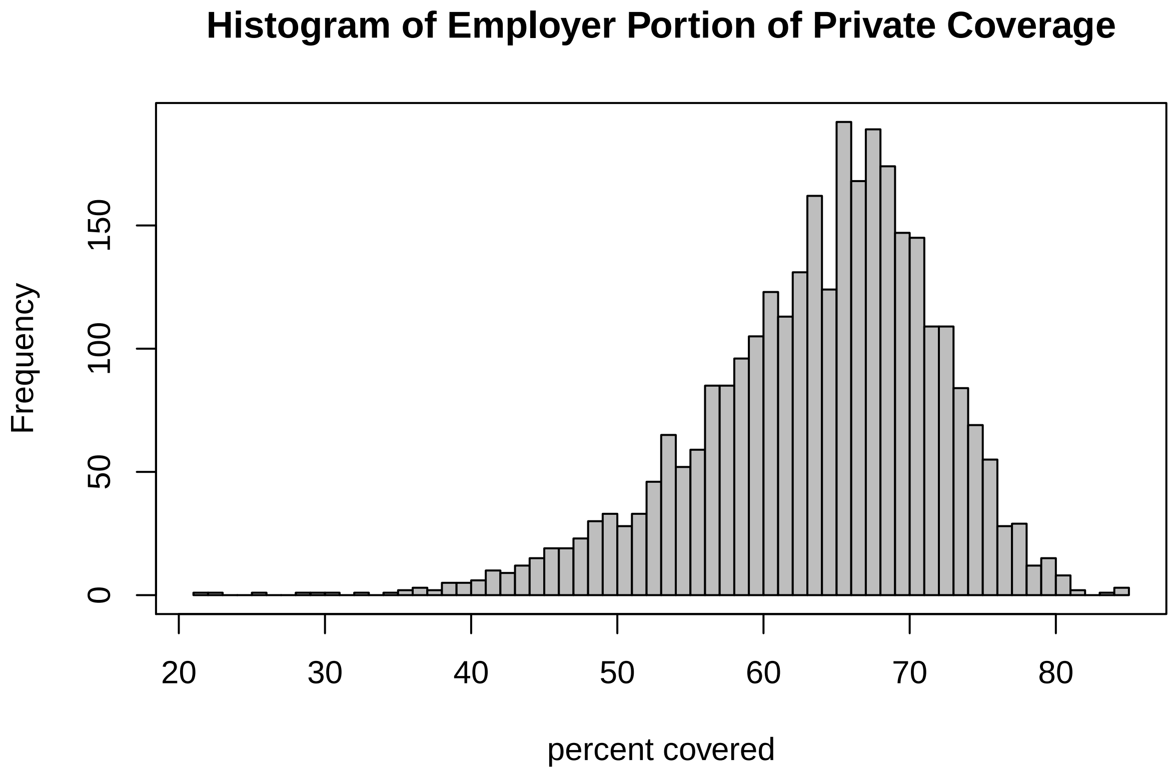
Employer-sponsored portion of the private coverage (EmpSponsoredPct variable)

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%.

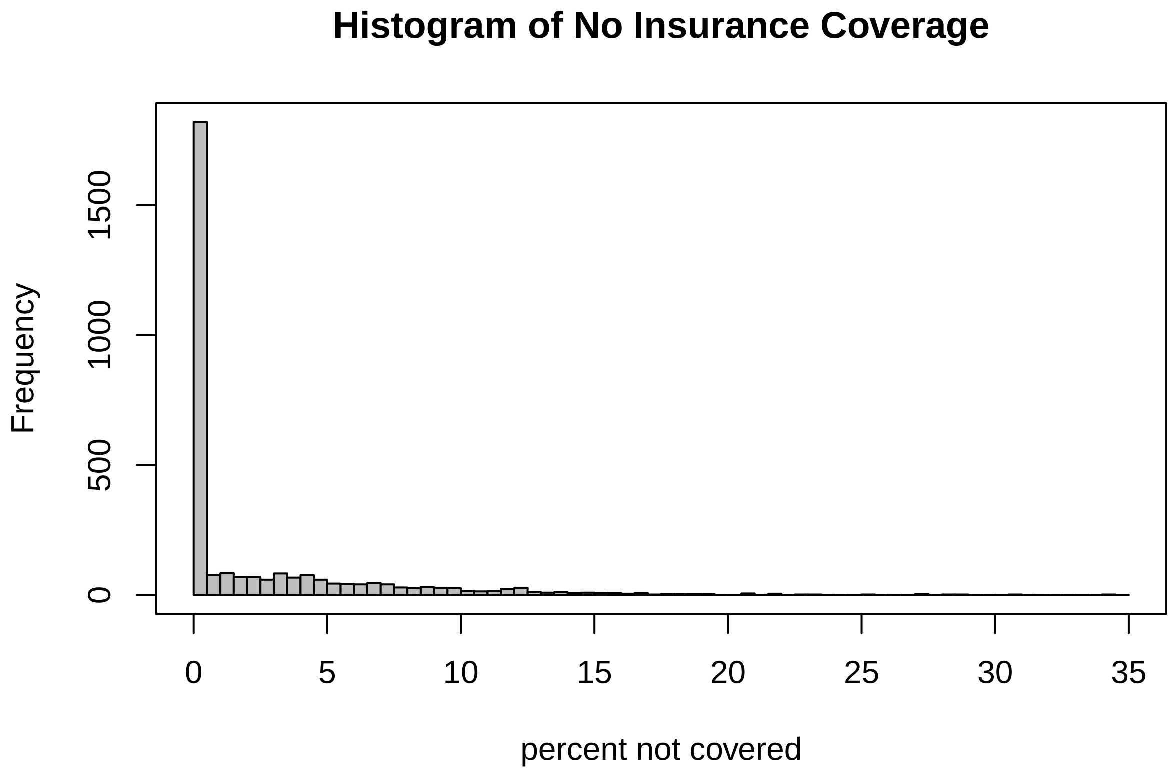
No insurance coverage (PctNoCoverage variable)

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.

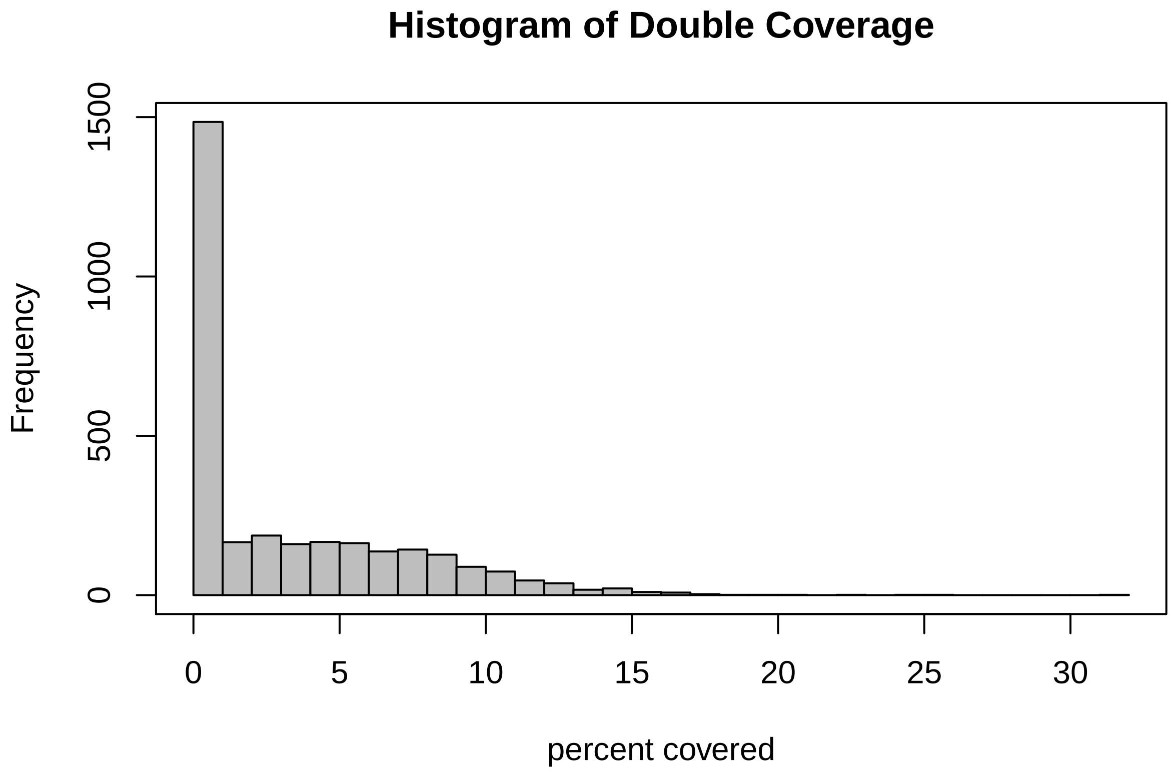
Coverage that includes both private and public components (PctDoubleCoverage variable)

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
```

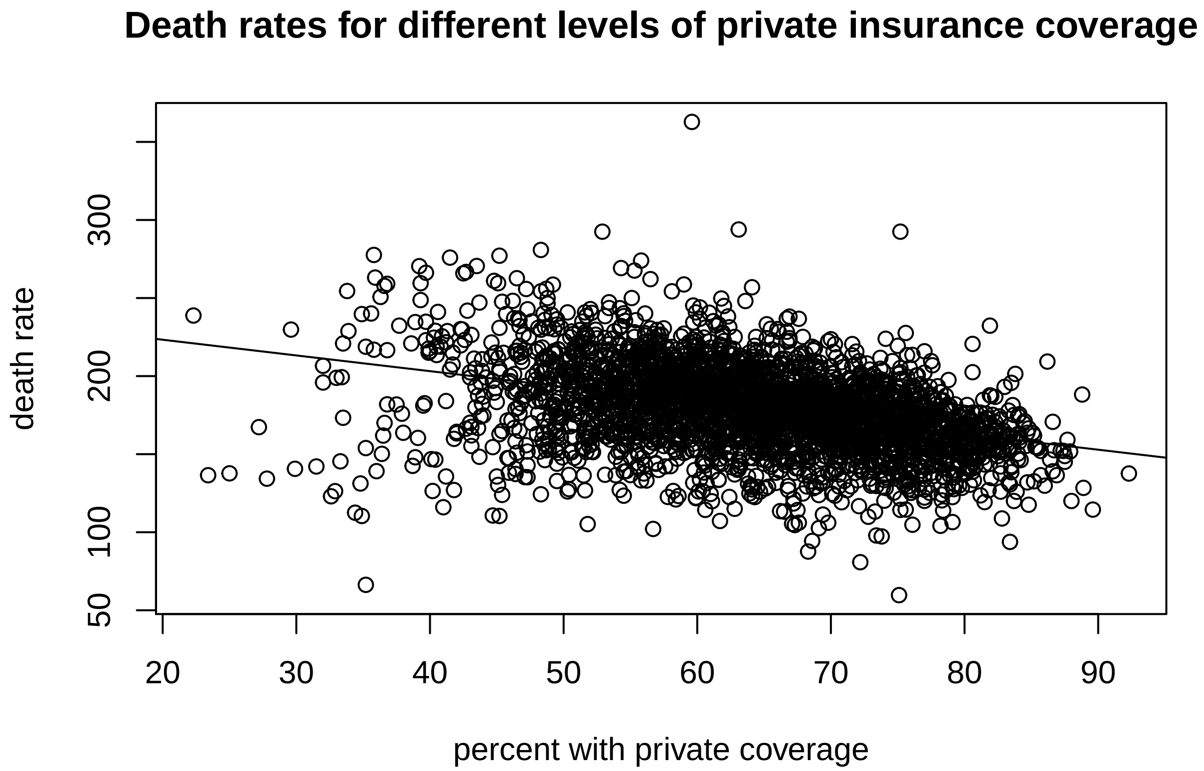
The result shows that 80% of the counties have less than 7% of the population with double health insurance. Therefore, similar to the previous case, its relative effect on the target variable will be minimal.

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)
```

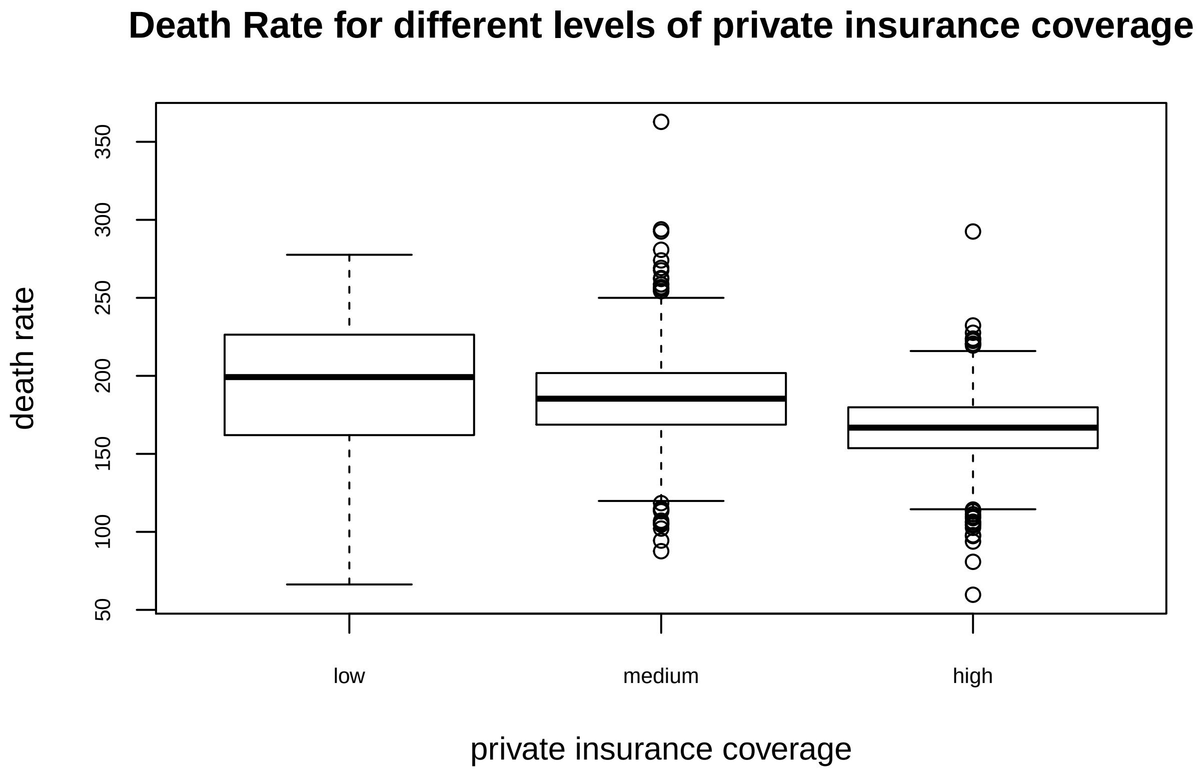
```
## [1] -0.3860655
```

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.

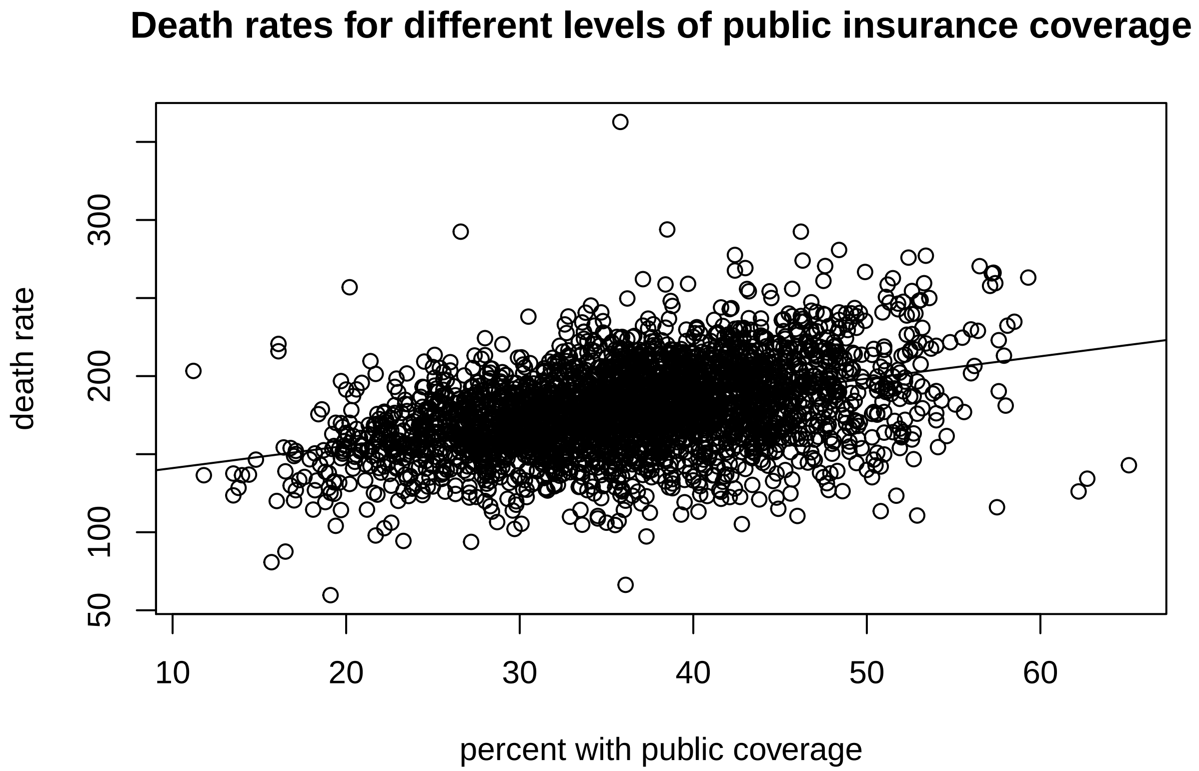
Summary of observations:

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:

```r
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))
```



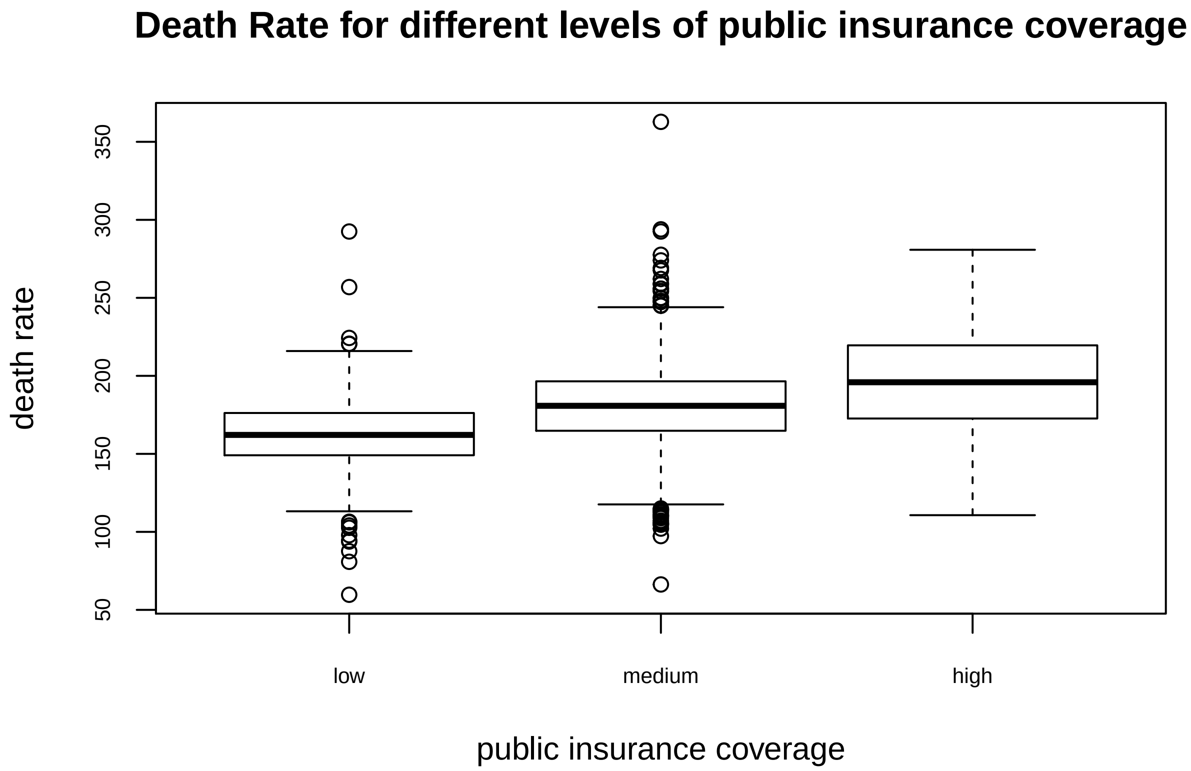
```
cor(cancer.df$deathRate, cancer.df$PctPublicCoverage)
```

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

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

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

```
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.

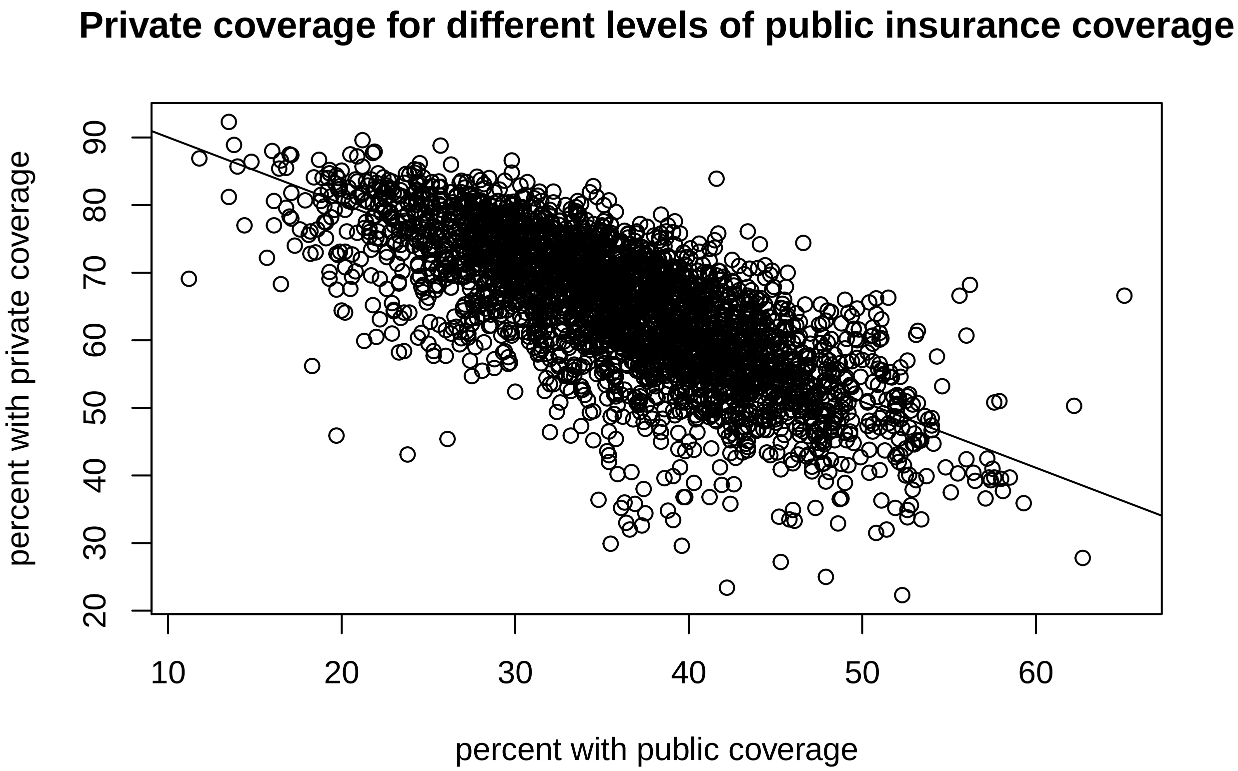
Summary of observations:

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:

```r
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))
```



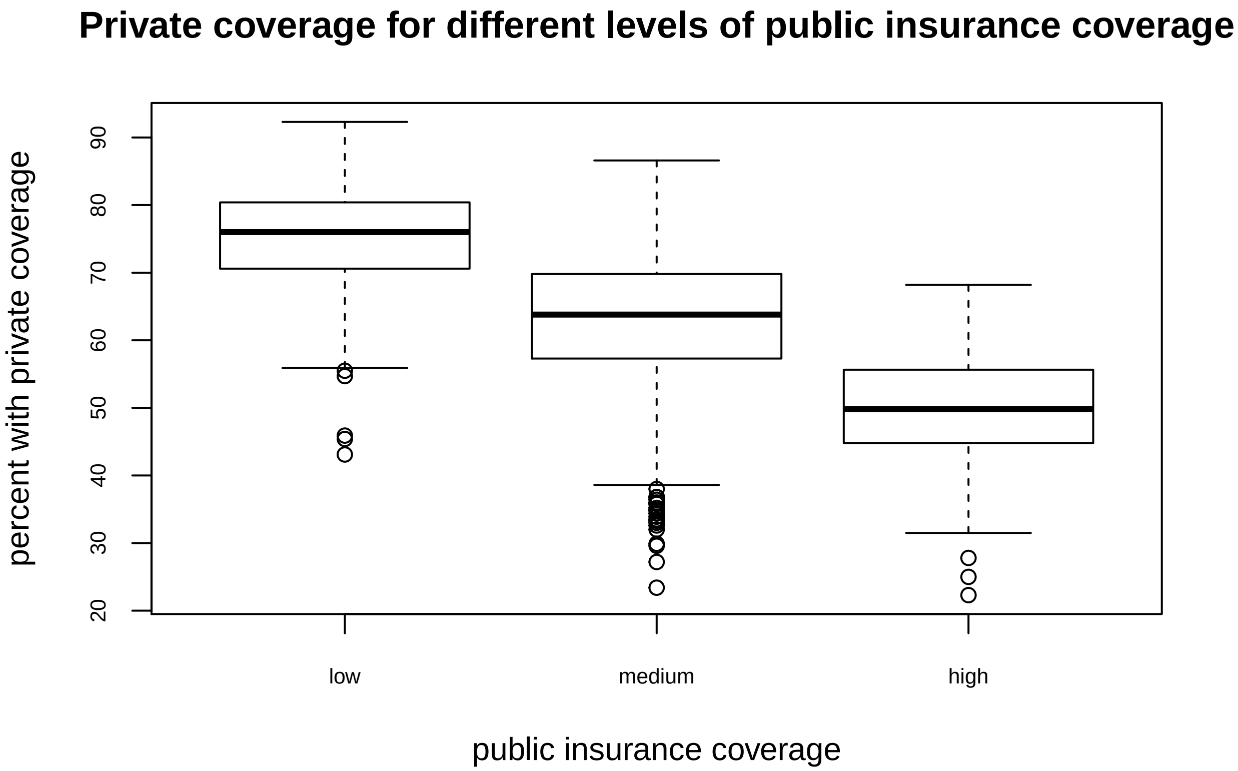
```
cor(cancer.df$PctPrivateCoverage, cancer.df$PctPublicCoverage)
```

```
## [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")
```



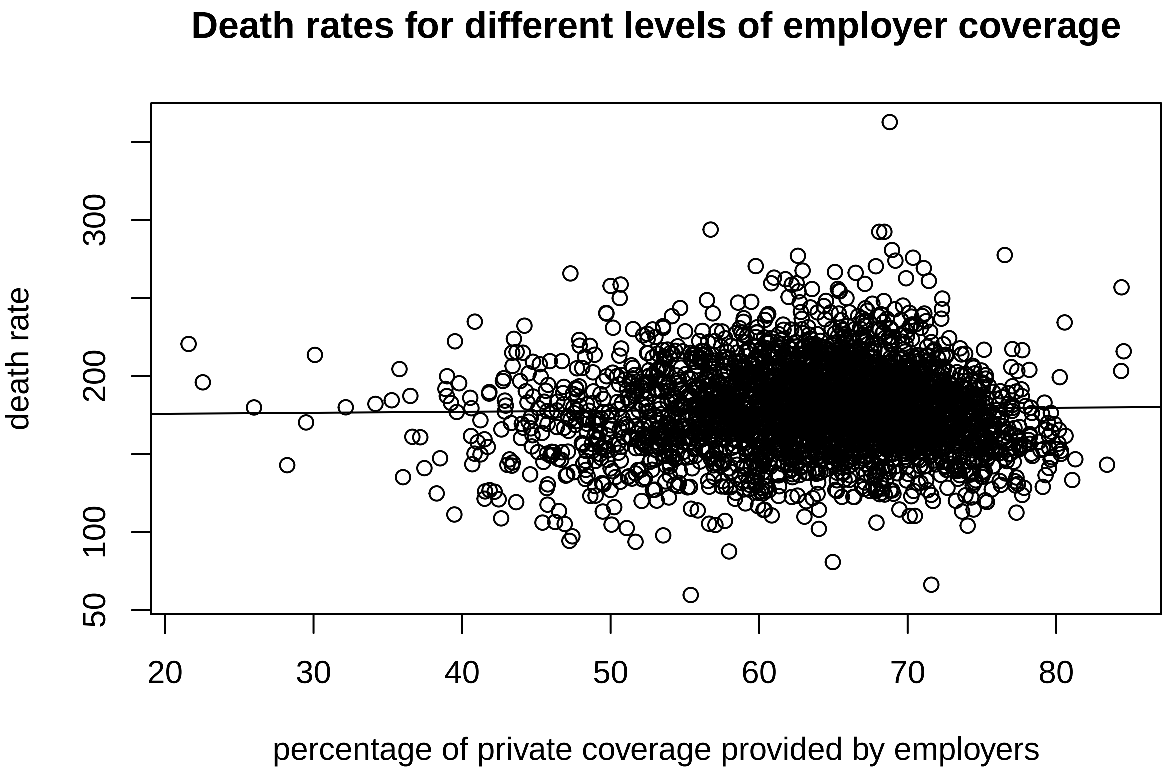
Summary of observations:

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))
```



```r
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.

Conclusion: insurance coverage per ce doesn't improve cancer mortality rates

1. The data analysis we performed has refuted our hypothesis that cancer patients who have access to health insurance have better chances of survival.
2. We also saw that private and public insurance demonstrate opposite relationships with cancer mortality rates.
3. One of the important findings is that higher levels of public insurance coverage are strongly correlated with lower percentage of private insurance coverage.
4. The relative amount of private coverage sponsored by employers have no detectable relationships with cancer mortality rates.
5. In order to explain these counter-intuitive results, as well as the 'threshold effect' of private coverage, we need to explore other variables that might directly influence these relationships.