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.

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 ...
```

```
summary(cancer.df)
```

```
##        X           avgAnnCount        medIncome        popEst2015
##  Min.   :   1.0   Min.   :    6.0   Min.   : 22640   Min.   :     827
##  1st Qu.: 762.5   1st Qu.:   76.0   1st Qu.: 38882   1st Qu.:   11684
##  Median :1524.0   Median :  171.0   Median : 45207   Median :   26643
##  Mean   :1524.0   Mean   :  606.3   Mean   : 47063   Mean   :  102637
##  3rd Qu.:2285.5   3rd Qu.:  518.0   3rd Qu.: 52492   3rd Qu.:   68671
##  Max.   :3047.0   Max.   :38150.0   Max.   :125635   Max.   :10170292
##
##  povertyPercent               binnedInc      MedianAge
##  Min.   : 3.20   (45201, 48021.6]  : 306   Min.   : 22.30
##  1st Qu.:12.15   (54545.6, 61494.5]: 306   1st Qu.: 37.70
##  Median :15.90   [22640, 34218.1]  : 306   Median : 41.00
##  Mean   :16.88   (42724.4, 45201]  : 305   Mean   : 45.27
##  3rd Qu.:20.40   (48021.6, 51046.4]: 305   3rd Qu.: 44.00
##  Max.   :47.40   (51046.4, 54545.6]: 305   Max.   :624.00
##                  (Other)           :1214
##  MedianAgeMale   MedianAgeFemale                            Geography
##  Min.   :22.40   Min.   :22.30   Abbeville County, South Carolina:   1
##  1st Qu.:36.35   1st Qu.:39.10   Acadia Parish, Louisiana        :   1
##  Median :39.60   Median :42.40   Accomack County, Virginia       :   1
##  Mean   :39.57   Mean   :42.15   Ada County, Idaho               :   1
##  3rd Qu.:42.50   3rd Qu.:45.30   Adair County, Iowa              :   1
##  Max.   :64.70   Max.   :65.70   Adair County, Kentucky          :   1
##                                  (Other)                         :3041
##  AvgHouseholdSize PercentMarried   PctNoHS18_24     PctHS18_24
##  Min.   :0.0221   Min.   :23.10   Min.   : 0.00   Min.   : 0.0
##  1st Qu.:2.3700   1st Qu.:47.75   1st Qu.:12.80   1st Qu.:29.2
##  Median :2.5000   Median :52.40   Median :17.10   Median :34.7
##  Mean   :2.4797   Mean   :51.77   Mean   :18.22   Mean   :35.0
##  3rd Qu.:2.6300   3rd Qu.:56.40   3rd Qu.:22.70   3rd Qu.:40.7
##  Max.   :3.9700   Max.   :72.50   Max.   :64.10   Max.   :72.5
##
##  PctSomeCol18_24 PctBachDeg18_24   PctHS25_Over   PctBachDeg25_Over
##  Min.   : 7.10   Min.   : 0.000   Min.   : 7.50   Min.   : 2.50
##  1st Qu.:34.00   1st Qu.: 3.100   1st Qu.:30.40   1st Qu.: 9.40
##  Median :40.40   Median : 5.400   Median :35.30   Median :12.30
##  Mean   :40.98   Mean   : 6.158   Mean   :34.80   Mean   :13.28
##  3rd Qu.:46.40   3rd Qu.: 8.200   3rd Qu.:39.65   3rd Qu.:16.10
##  Max.   :79.00   Max.   :51.800   Max.   :54.80   Max.   :42.20
##  NA's   :2285
##  PctEmployed16_Over PctUnemployed16_Over PctPrivateCoverage
##  Min.   :17.60      Min.   : 0.400       Min.   :22.30
##  1st Qu.:48.60      1st Qu.: 5.500       1st Qu.:57.20
##  Median :54.50      Median : 7.600       Median :65.10
```

```
##  Mean   :54.15      Mean   : 7.852       Mean   :64.35
##  3rd Qu.:60.30      3rd Qu.: 9.700       3rd Qu.:72.10
##  Max.   :80.10      Max.   :29.400       Max.   :92.30
##  NA's   :152
##  PctEmpPrivCoverage PctPublicCoverage    PctWhite         PctBlack
##  Min.   :13.5       Min.   :11.20     Min.   : 10.20   Min.   : 0.0000
##  1st Qu.:34.5       1st Qu.:30.90     1st Qu.: 77.30   1st Qu.: 0.6207
##  Median :41.1       Median :36.30     Median : 90.06   Median : 2.2476
##  Mean   :41.2       Mean   :36.25     Mean   : 83.65   Mean   : 9.1080
##  3rd Qu.:47.7       3rd Qu.:41.55     3rd Qu.: 95.45   3rd Qu.:10.5097
##  Max.   :70.7       Max.   :65.10     Max.   :100.00   Max.   :85.9478
##
##     PctAsian        PctOtherRace     PctMarriedHouseholds   BirthRate
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :22.99        Min.   : 0.000
##  1st Qu.: 0.2542   1st Qu.: 0.2952   1st Qu.:47.76        1st Qu.: 4.521
##  Median : 0.5498   Median : 0.8262   Median :51.67        Median : 5.381
##  Mean   : 1.2540   Mean   : 1.9835   Mean   :51.24        Mean   : 5.640
##  3rd Qu.: 1.2210   3rd Qu.: 2.1780   3rd Qu.:55.40        3rd Qu.: 6.494
##  Max.   :42.6194   Max.   :41.9303   Max.   :78.08        Max.   :21.326
##
##    deathRate
##  Min.   : 59.7
##  1st Qu.:161.2
##  Median :178.1
##  Mean   :178.7
##  3rd Qu.:195.2
##  Max.   :362.8
##
```

Validation of Independent Variables

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

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. We set these outliers to NA 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
```

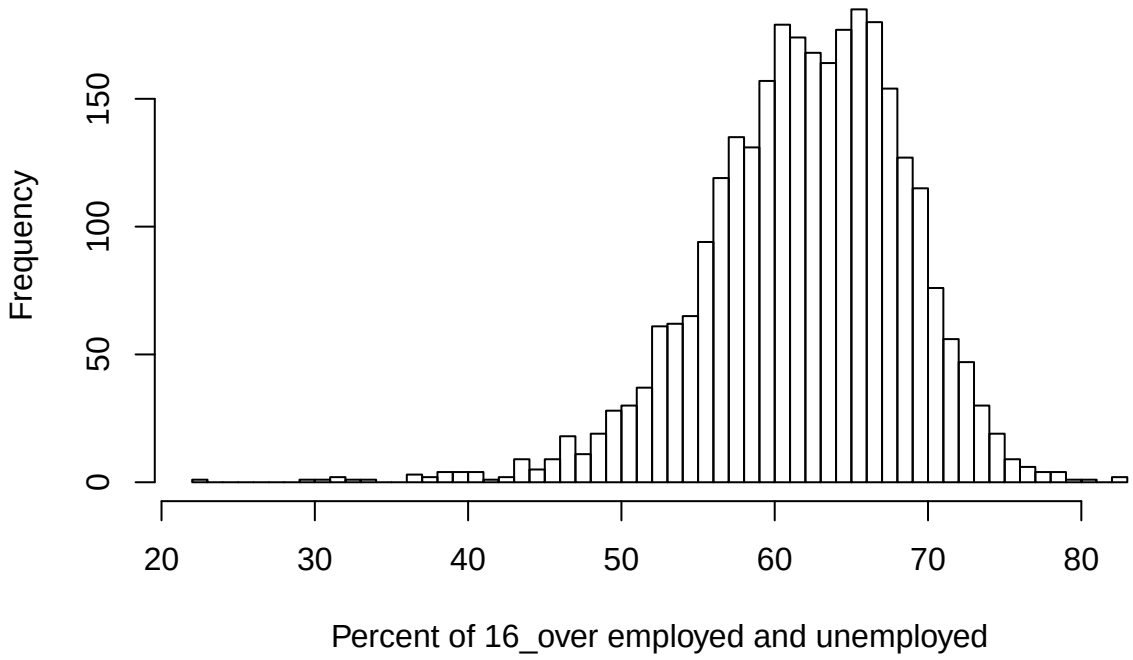
```
cancer.df$MedianAge[cancer.df$MedianAge>300]<-NA  # Set outlier values to NA
```

The sum of the variables percentage employed and unemployed over 16 has a surprisingly broad distribution around the mean of 62.01, when one would expect it to be close to (if not) 100%.

```
Emp.UnEmp<-cancer.df$PctEmployed16_Over+cancer.df$PctUnemployed16_Over
summary(Emp.UnEmp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   22.40   58.00   62.50   62.01   66.60   82.70     152
```

```
hist(Emp.UnEmp,breaks="fd",main="Distribution of Employment data per county",xlab="Percent of 16_over em
```



We note this for now as we analyze other variables.

Validation of potential Outcome Variables.

There are 2 variables we considered as potential outcome variables: Death Rate (assumed to be number of death per 100,000 population) and Average Annual Count (assumed to be annual cancer.df Indidence Rate)

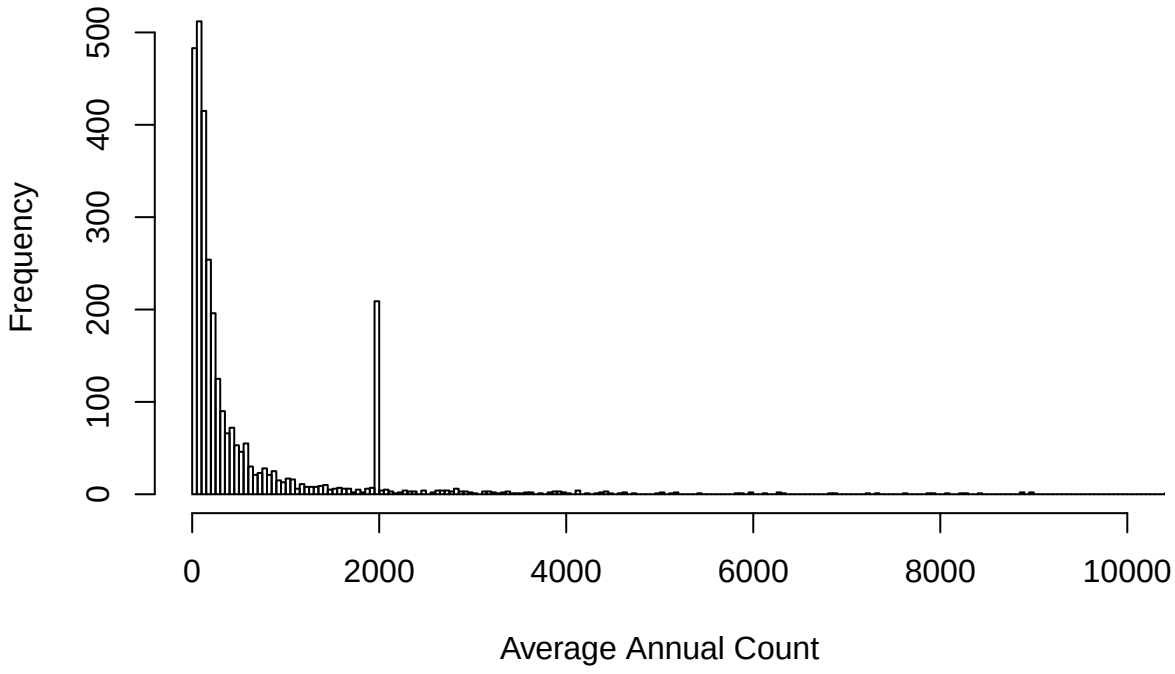
```
#Annual Indident Rate is better expressed as a percentage of county population

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
```

```
#Look for where the outlier might be coming from

hist(cancer.df$avgAnnCount,breaks="fd",main="Average Annual Count Distrubution",xlab="Average Annual Cou
```




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

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

```
#Clearly all of these have the exact same erroneous value for Average Annual Count.
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

summary(cancer.df$avgAnnCount)
```

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

```
cancer.df$AnnCountPercent<-with(cancer.df,100*avgAnnCount/popEst2015) #Recalculate percentages
summary(cancer.df$AnnCountPercent)
```

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

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

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

Analysis of Key Variables

We have chosen to focus on the following variables in this study

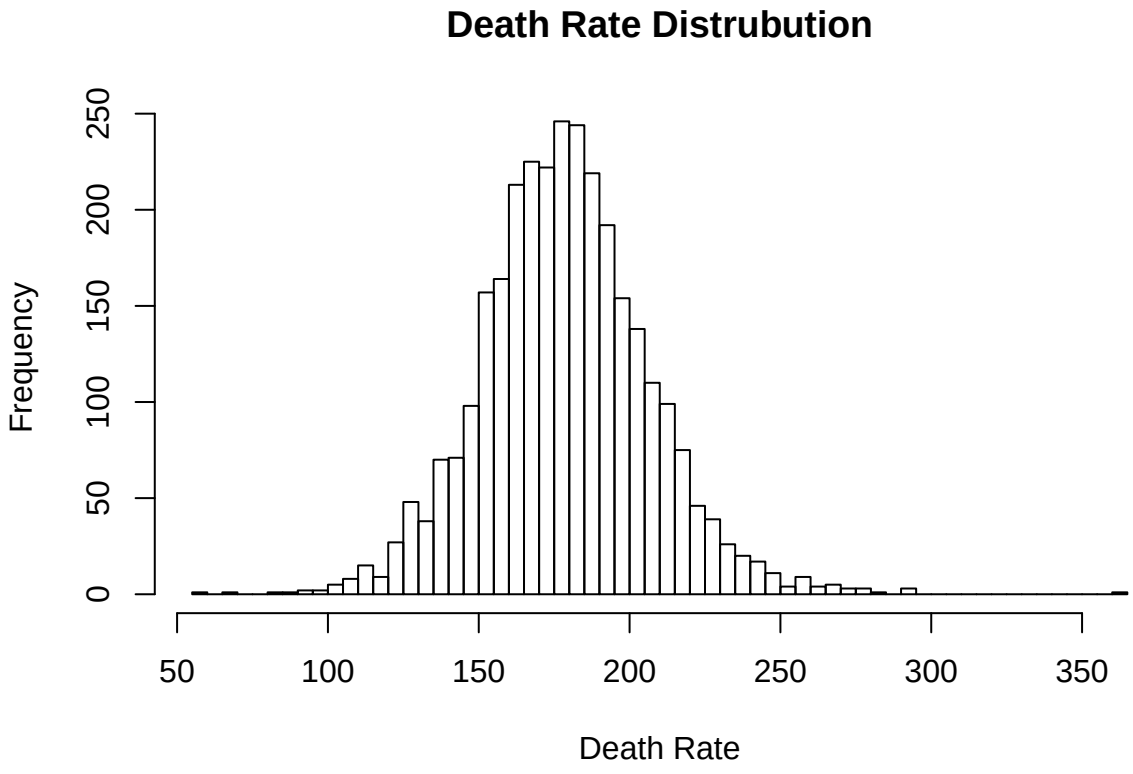
Table 1: Key Variables

Variable Name	Description
DeathRate	Our output variable
avgAnnCount	2009-2013 mean incidences per county
AnnCountPercent	2009-2013 Incident Rate expressed as a percent of population
popEst2015	Estimated population by county 2015
PctPrivateCoverage	Percentage of the population with private insurance coverage
PctPublicCoverage	Percentage of the population with public insurance coverage
PctEmpPrivCoverage	Percentage of population with empoyer private insurance coverage
povertyPercent	Percent of population below poverty line
MedianAge	Median population age
medIncome	Median Income

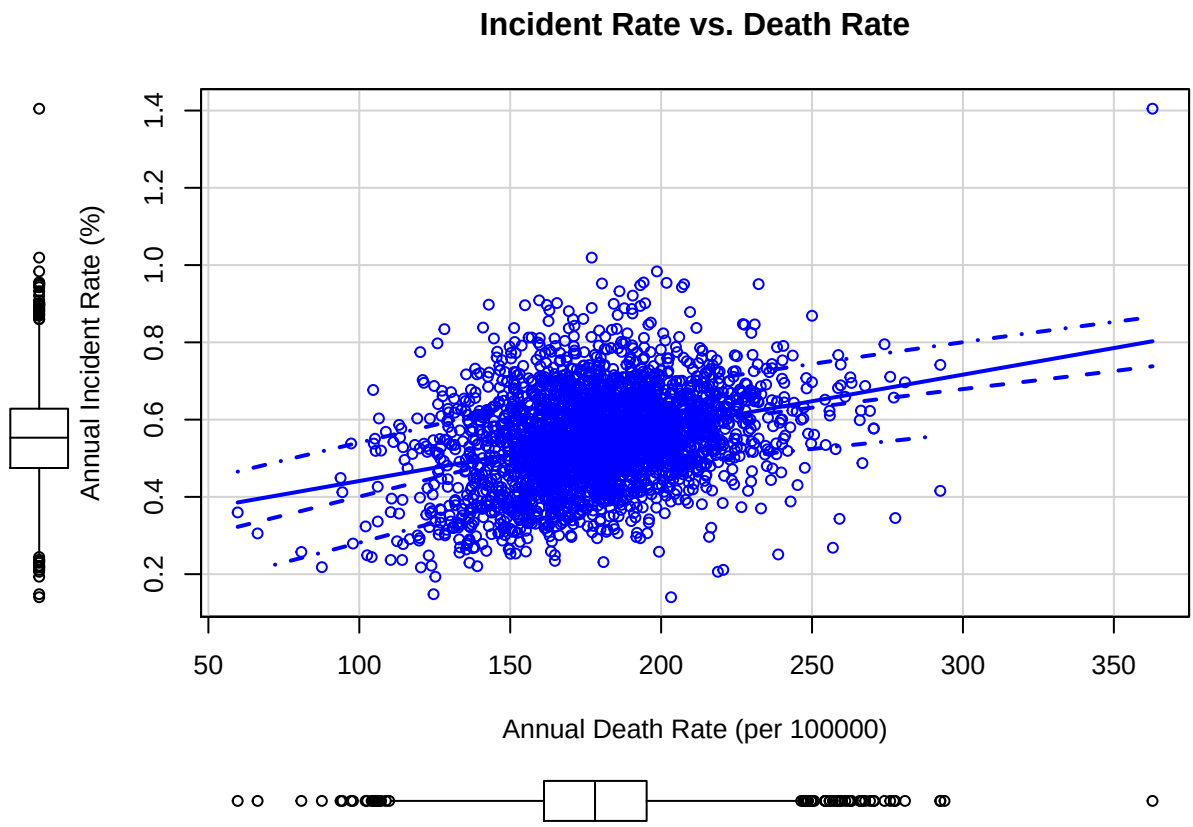
```
cancer.df[cancer.df$deathRate > 300,]
```

```
##         X avgAnnCount medIncome popEst2015 povertyPercent
## 1490 1490         214     40207      15234           24.3
##               binnedInc MedianAge MedianAgeMale MedianAgeFemale
## 1490 (37413.8, 40362.7]      40.3          42.3            36.9
##                  Geography AvgHouseholdSize PercentMarried PctNoHS18_24
## 1490 Union County, Florida             2.58           36.4           27
##      PctHS18_24 PctSomeCol18_24 PctBachDeg18_24 PctHS25_Over
## 1490       45.1              NA               0         37.4
##      PctBachDeg25_Over PctEmployed16_Over PctUnemployed16_Over
## 1490               5.5                 NA                 11.7
##      PctPrivateCoverage PctEmpPrivCoverage PctPublicCoverage PctWhite
## 1490               59.6                 41              35.8 73.96485
##      PctBlack  PctAsian PctOtherRace PctMarriedHouseholds BirthRate
## 1490 21.59173 0.6451188     1.533803             50.01288  3.739774
##      deathRate AnnCountPercent
## 1490     362.8        1.404753
```

```
hist(cancer.df$deathRate,breaks="fd",main="Death Rate Distrubution",xlab="Death Rate")
```




```
scatterplot(AnnCountPercent~deathRate,ylab="Annual Incident Rate (%)", xlab="Annual Death Rate (per 100
```



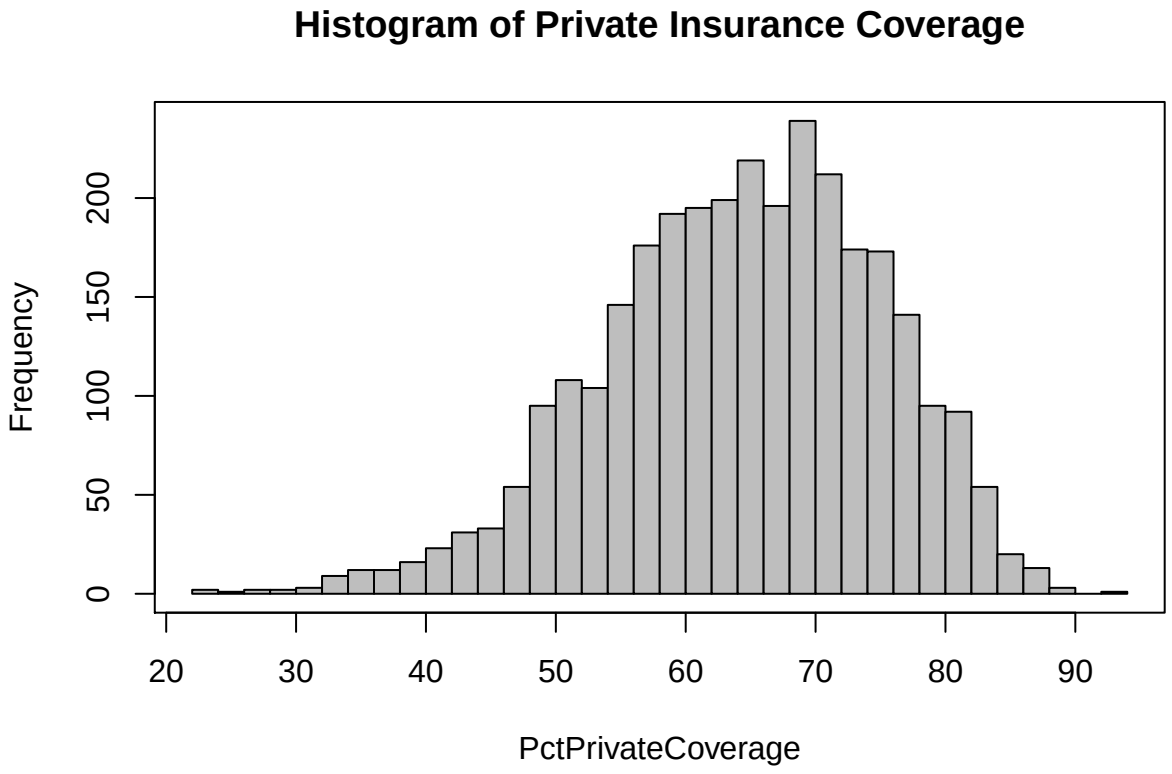
We see we still have a single outlier with Incident Rate AND Death Rate being much higher than all other counties.

PctPrivateCoverage

```
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 Insuranc
box()
```

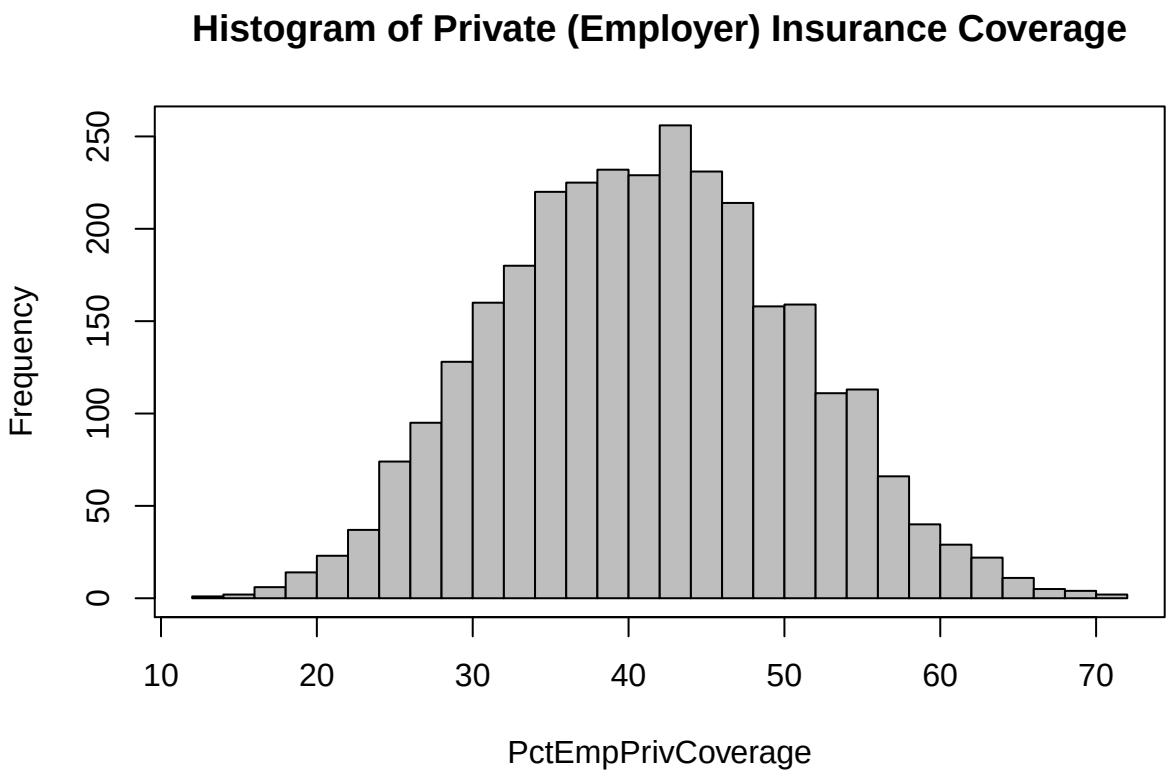


PctEmpPrivCoverage

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    13.5    34.5    41.1    41.2    47.7    70.7
```

```
with(cancer.df, hist(PctEmpPrivCoverage  , breaks="FD", col = "gray", main="Histogram of Private (Emplo
box()
```

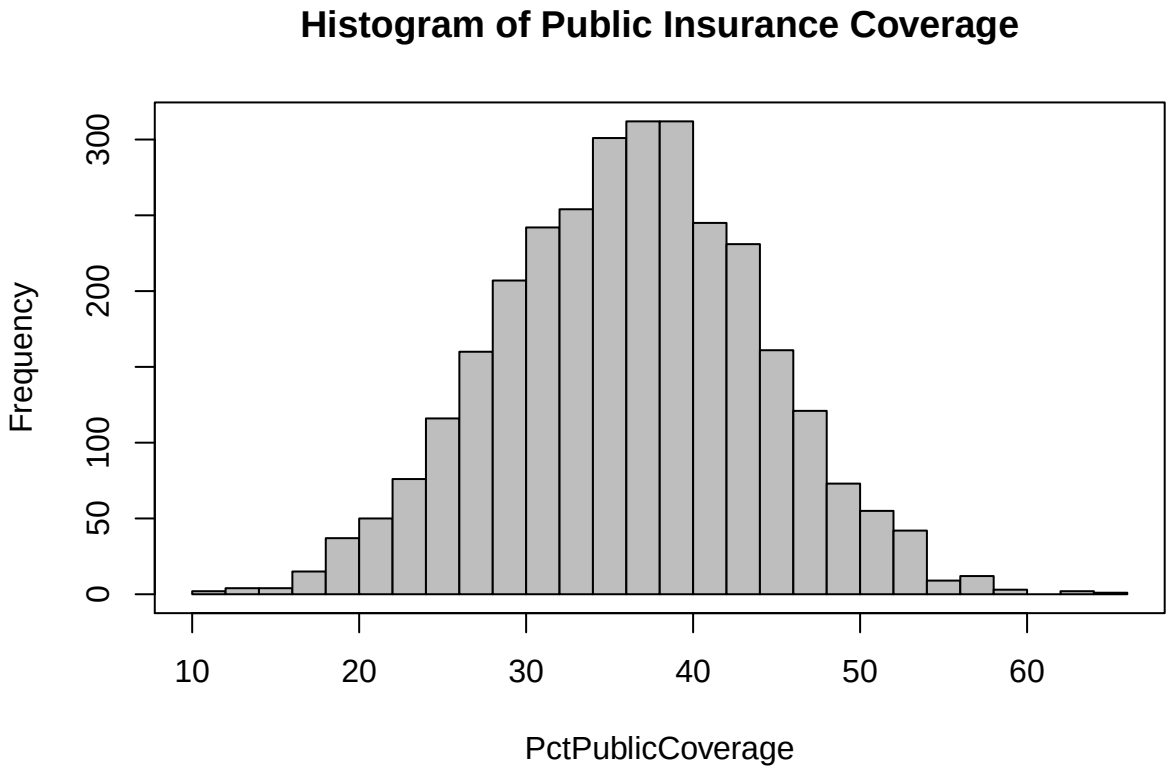


PctPublicCoverage

```
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 C
box()
```

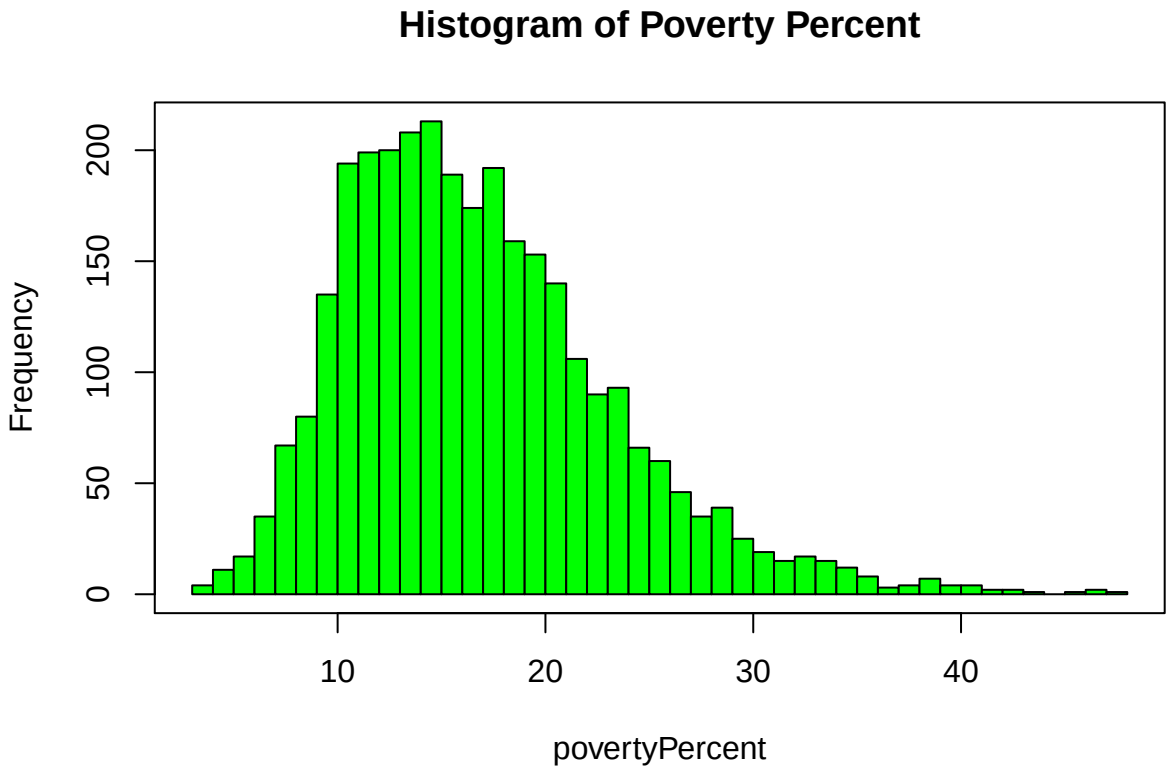


povertyPercent

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.20   12.15   15.90   16.88   20.40   47.40
```

```
with(cancer.df, hist(povertyPercent, breaks="FD", col = "green", main="Histogram of Poverty Percent"))
box()
```

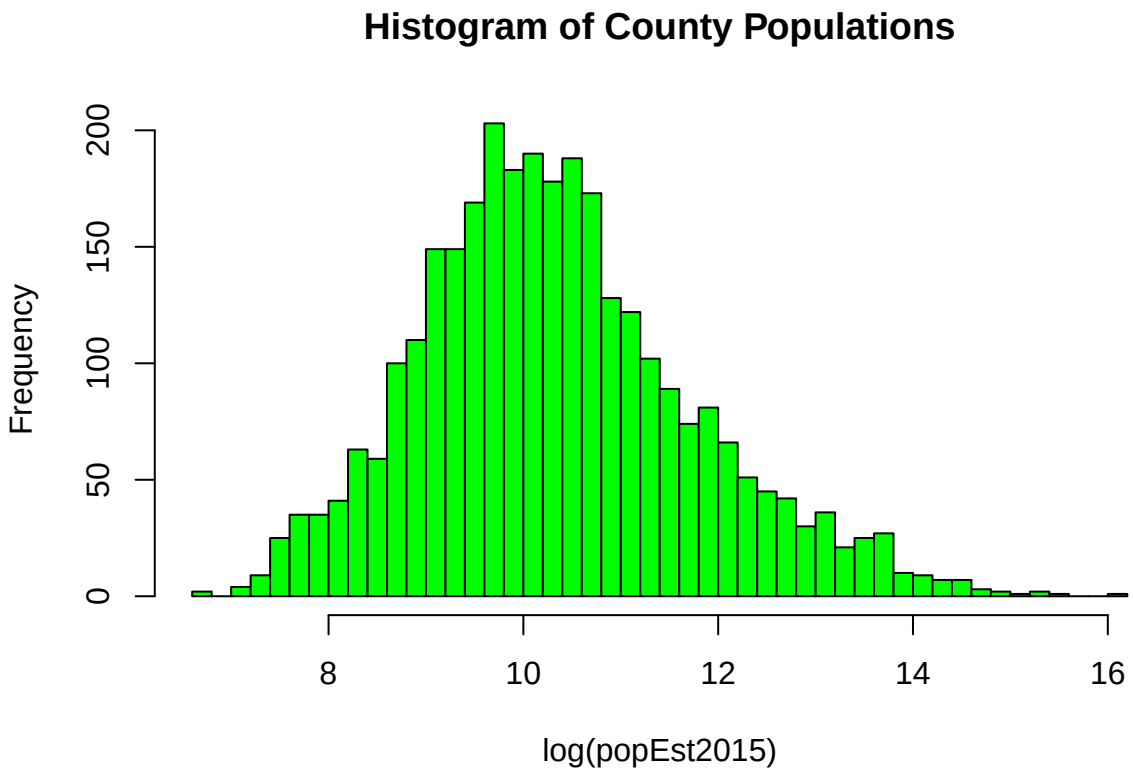


popEst2015

```
summary(cancer.df$popEst2015)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##       827     11680     26640    102600     68670 10170000
```

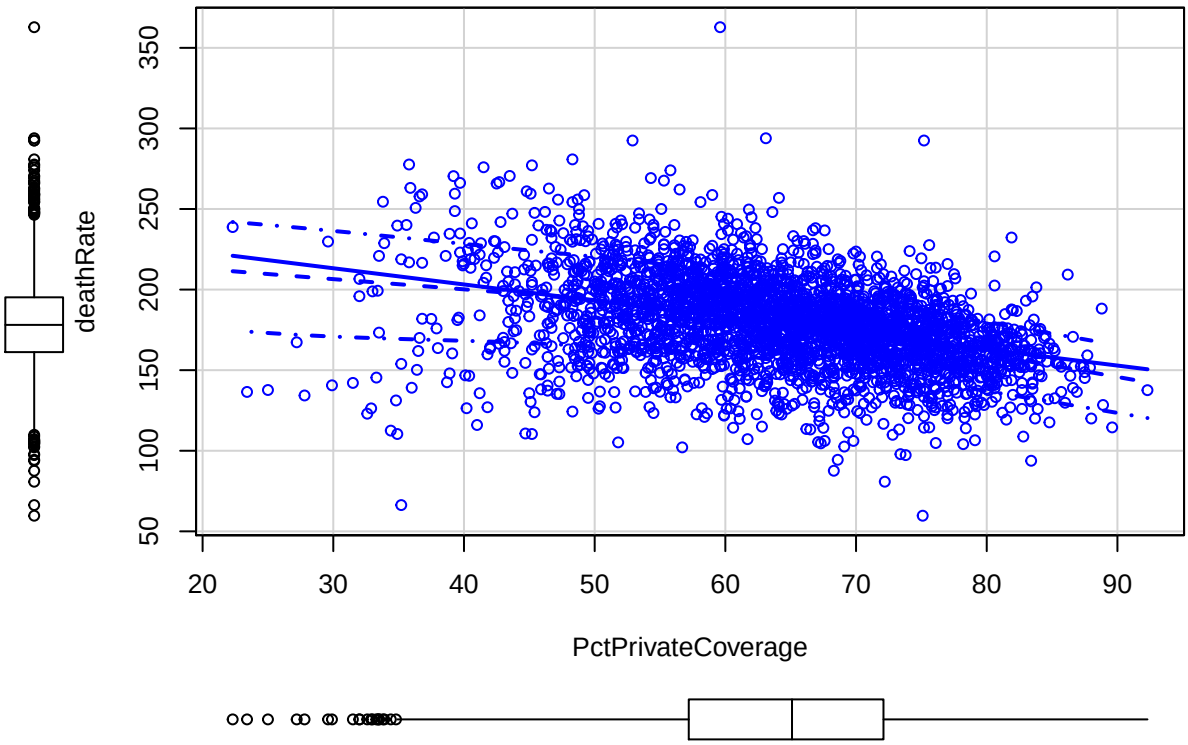
```
with(cancer.df, hist(log(popEst2015), breaks="FD", col = "green", main="Histogram of County Populations
```



Analysis of Key Relationships

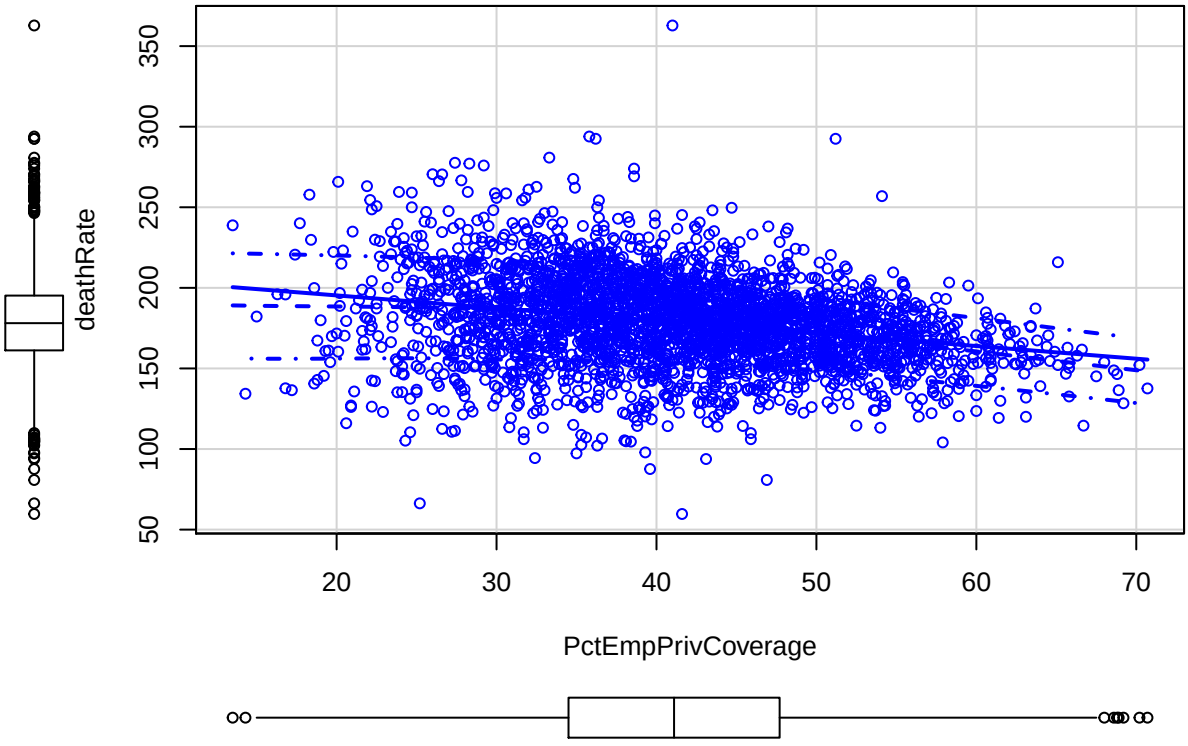
Death Rate vs. Kind of Coverage

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



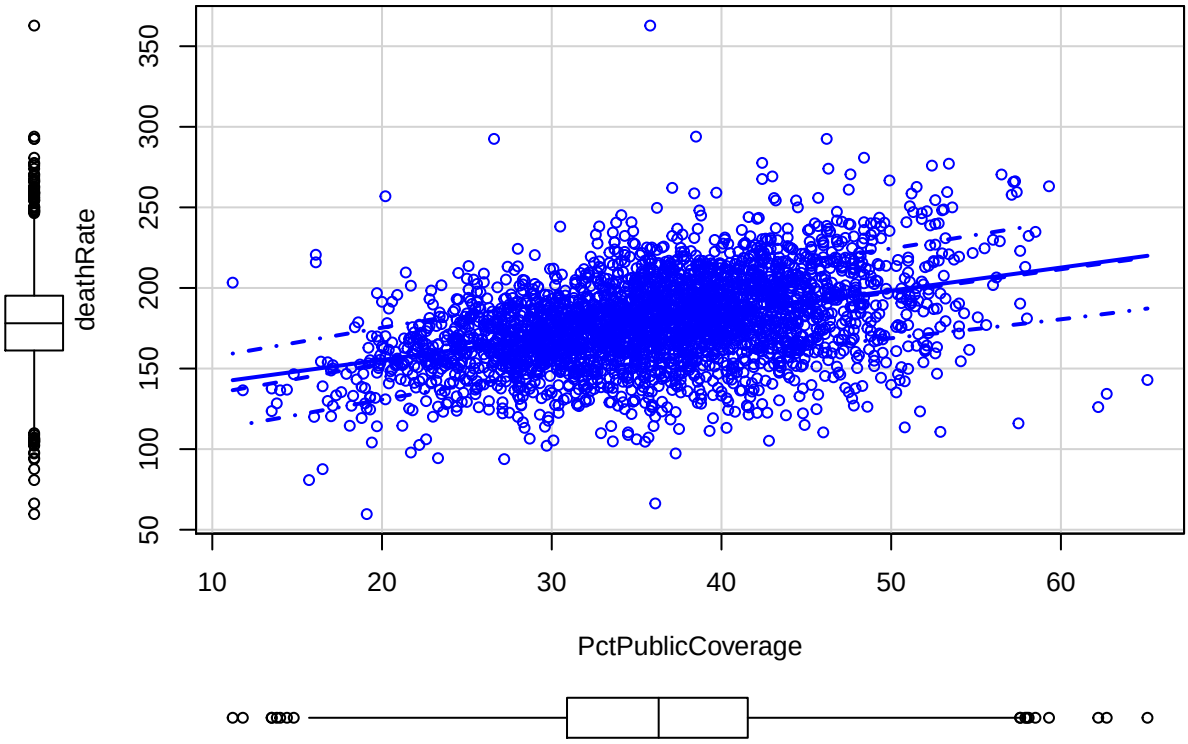
There clearly is a negative correlation between Mortality Rate and the percent of the population that has private coverage. This is comparable when looking at percent of population with employer sponsored private coverage as seen in the next plot.

```
## [1] -0.2673994
```

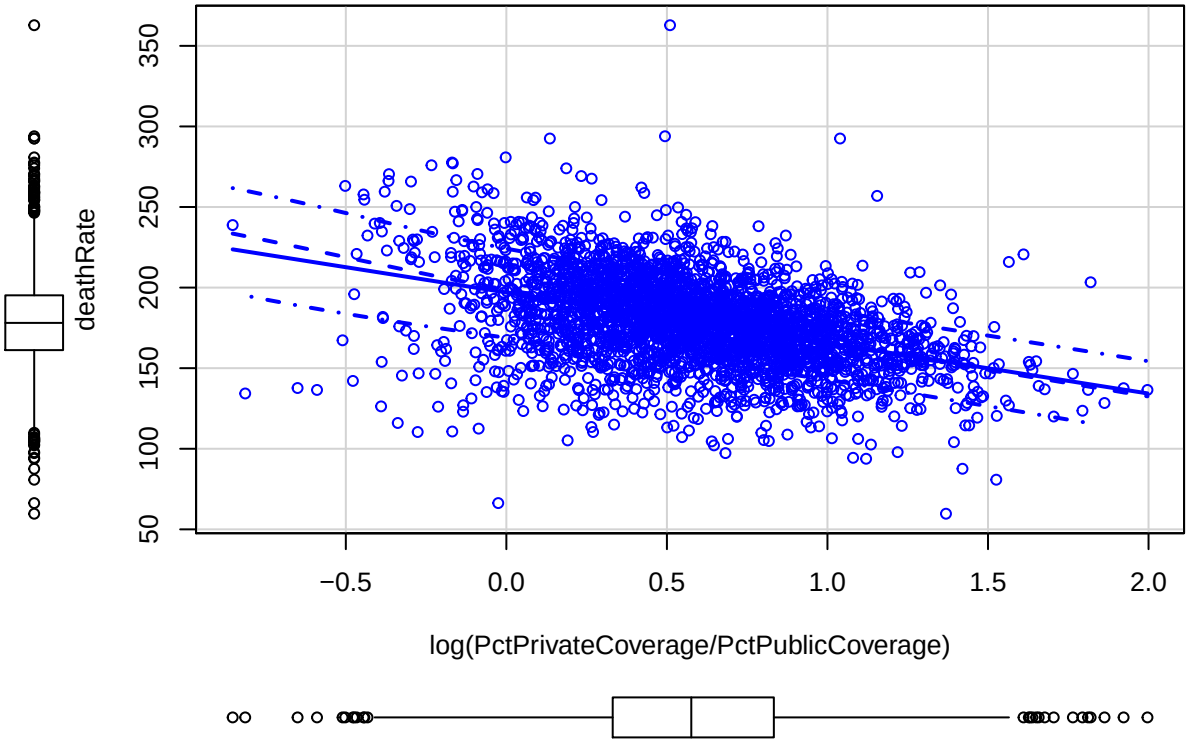


This intuitively suggests that percent of public coverage should have a positive correlation with Mortality Rate.

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

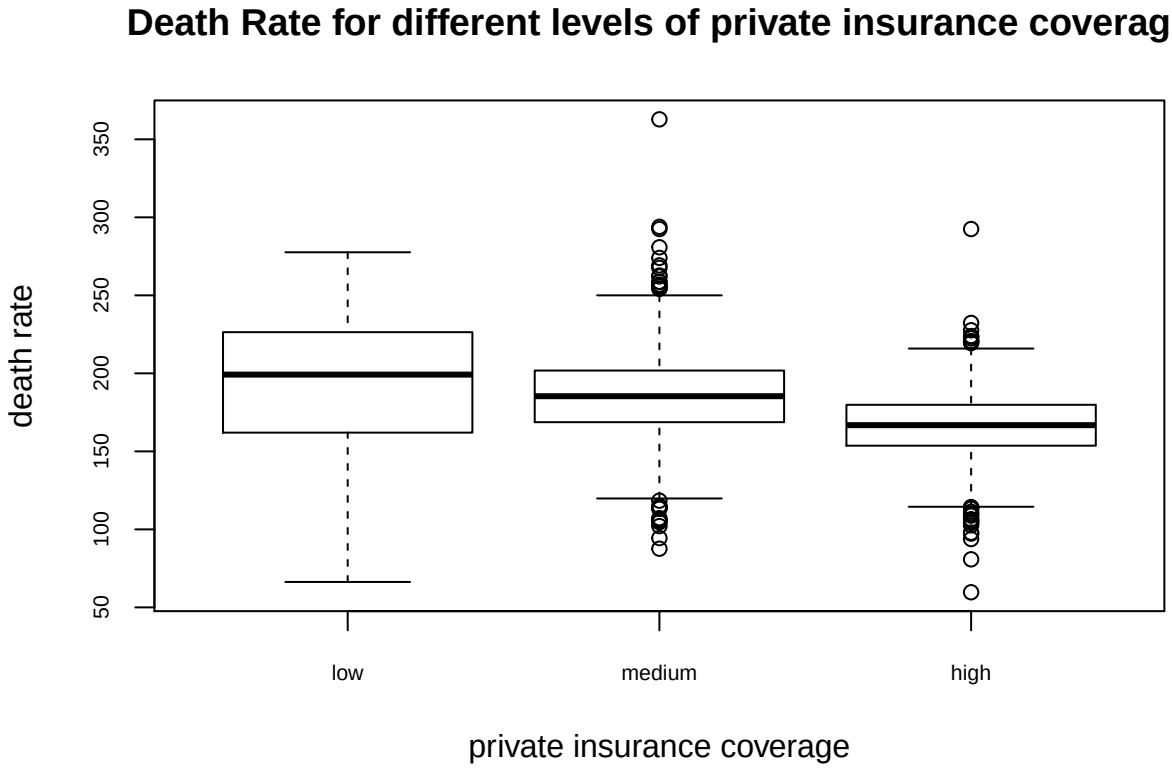



```
## [1] -0.4232724
```

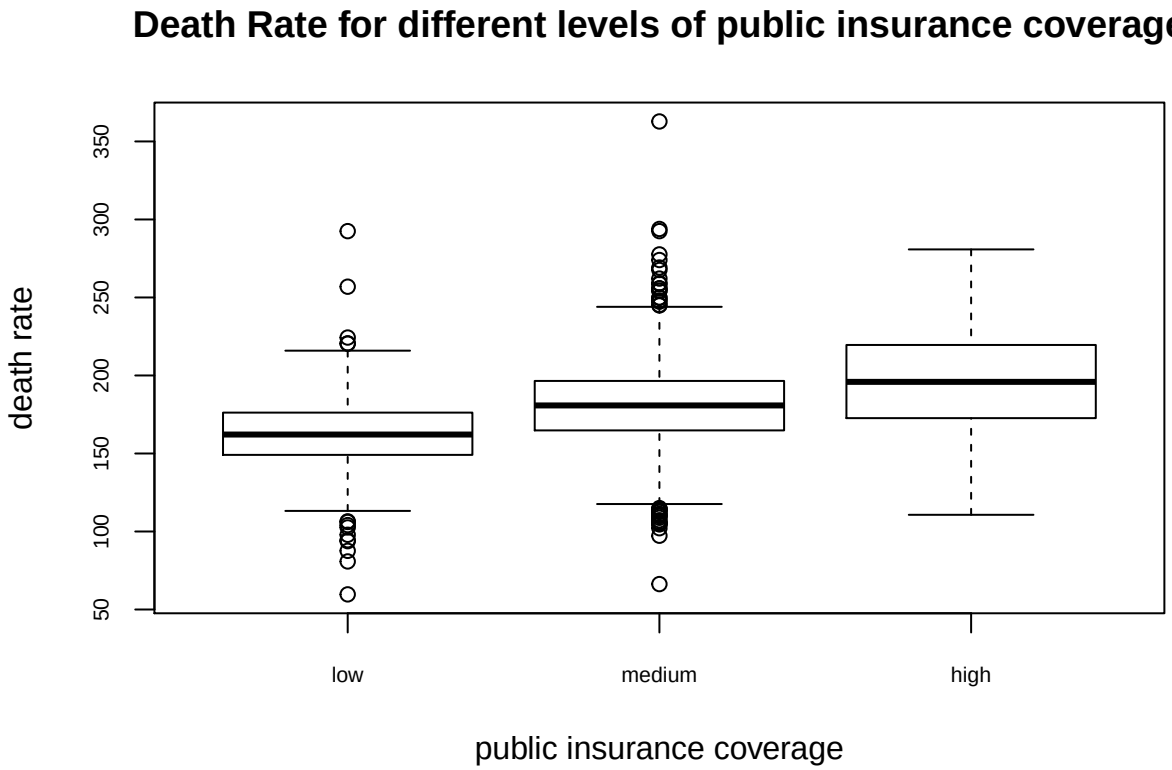


An alternate way of looking at this is by grouping into "levels of coverage".. This clearly shows the strong correlation between higher percent of public coverage and mortality rate.

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



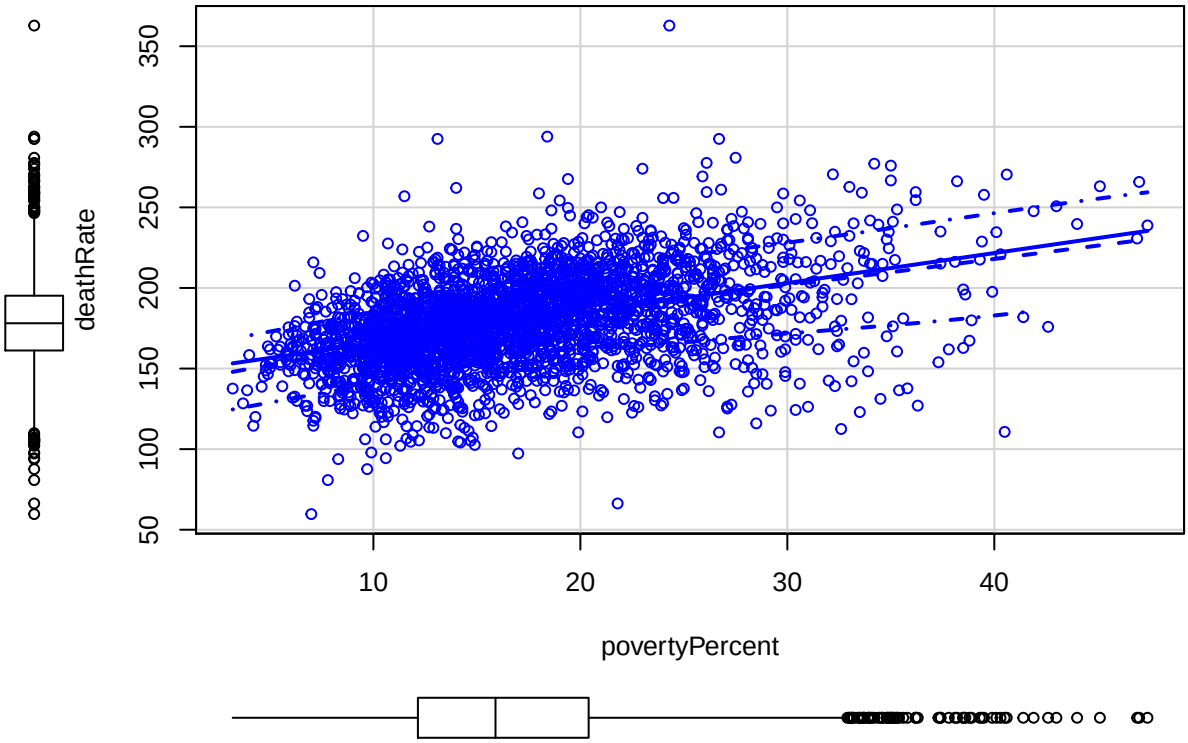

```
## [1] "[11.1,29.2]" "(29.2,47.1]" "(47.1,65.2]"
```



Death Rate vs. Percent under Poverty line

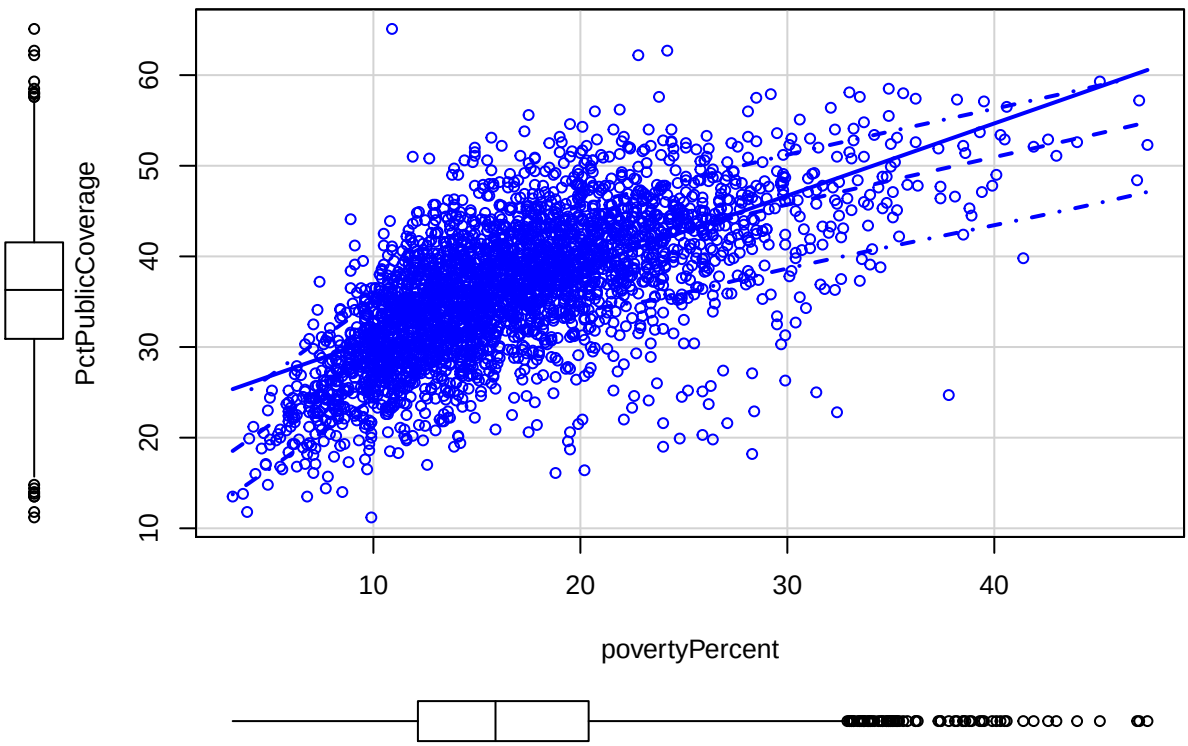
There is clearly a strong correlation between prevalence of poverty and mortality rte.

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



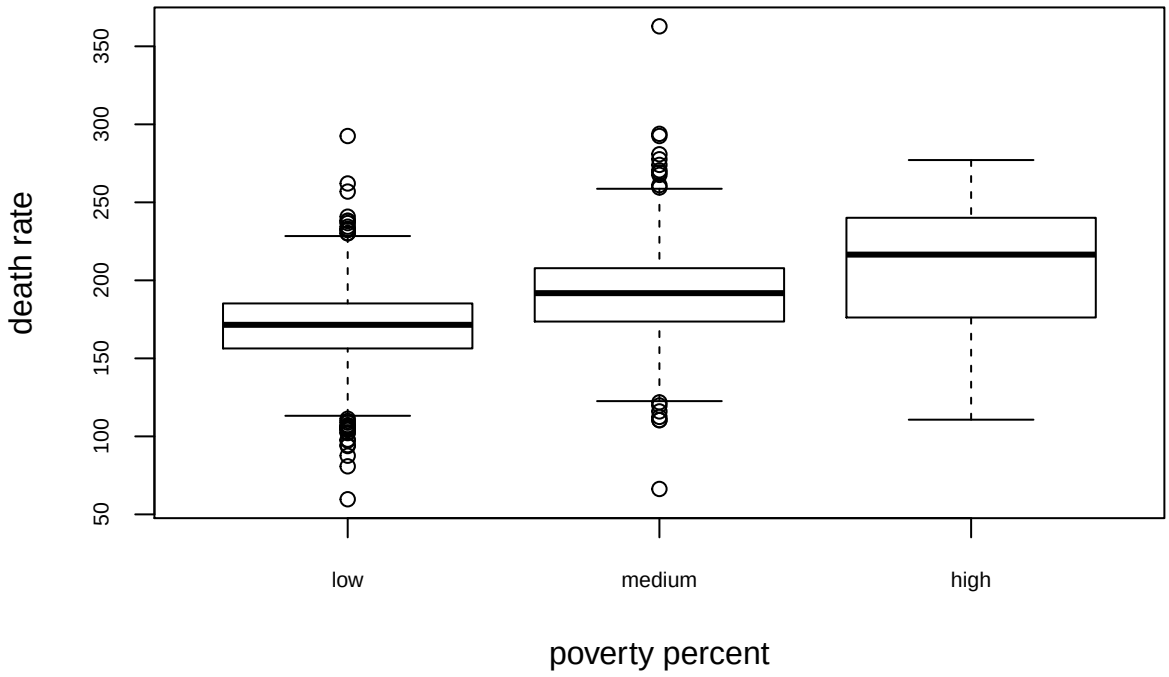
In comparing prevalance of poverty to public coverage rates though, one can see a strong corelation as one woudl expect

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

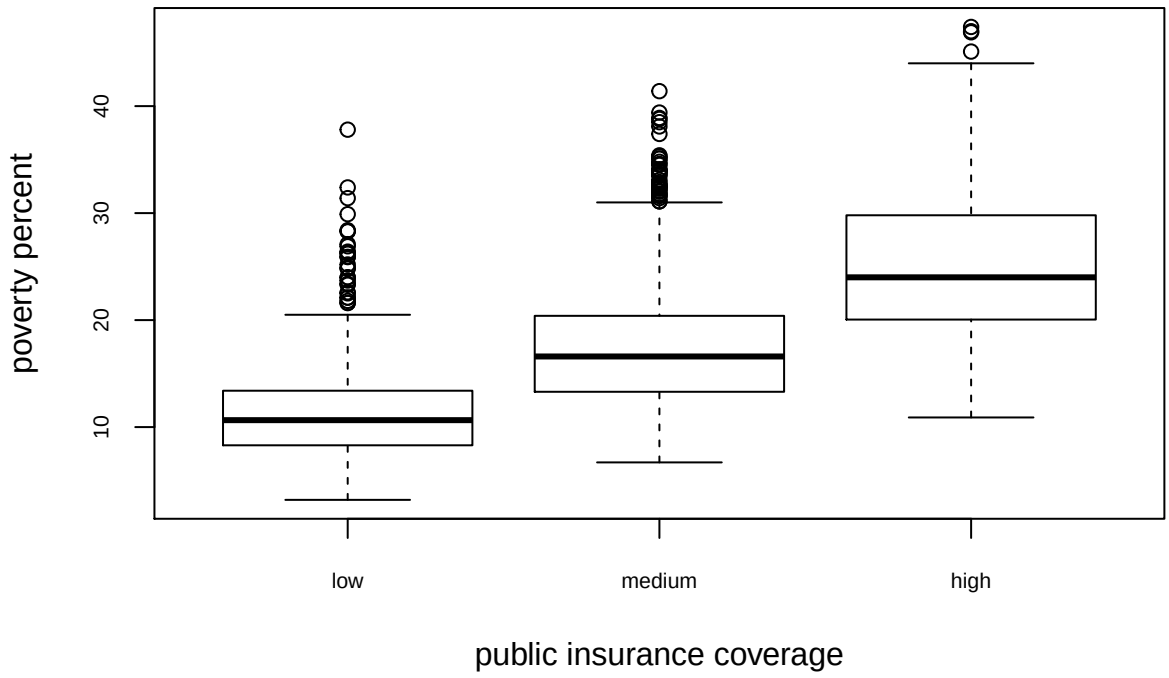



```
## [1] "[3.16,17.9]" "(17.9,32.7]" "(32.7,47.4]"
```

Death Rate for different levels of poverty percent



Poverty percent for different levels of public insurance coverage



***

Analysis of Secondary Effects

What secondary variables might have confounding effects on the relationships you have identified? Ex- plain how these variables affect your understanding of the data.

Conclusion

Summarize your exploratory analysis. What can you conclude based on your analysis?