Cancer Mortality Exploration

w203 Teaching Team

Background

In this lab, imagine that your team is hired by a health government agency. They would like to understand factors that predict cancer mortality rates, with the ultimate aim of identifying communities for social interventions, and of understanding which interventions are likely to have the most impact. Your team was hired to perform an exploratory analysis to help the agency address their goals.

Data

You are given a dataset for a selection of US counties, **"cancer.csv"**. The dependent (or target) variable in this data is named "deathRate".

The labels of some of the variables are listed below; the rest of the variables should be self-explanatory.

```
# Load CSV data
cancer = read.csv("cancer.csv")
```

Introduction

```
# Show variable names
colnames(cancer)
```

```
##  [1] "X"                    "avgAnnCount"          "medIncome"
##  [4] "popEst2015"           "povertyPercent"       "binnedInc"
##  [7] "MedianAge"            "MedianAgeMale"        "MedianAgeFemale"
## [10] "Geography"            "AvgHouseholdSize"     "PercentMarried"
## [13] "PctNoHS18_24"         "PctHS18_24"           "PctSomeCol18_24"
## [16] "PctBachDeg18_24"      "PctHS25_Over"         "PctBachDeg25_Over"
## [19] "PctEmployed16_Over"   "PctUnemployed16_Over" "PctPrivateCoverage"
## [22] "PctEmpPrivCoverage"   "PctPublicCoverage"    "PctWhite"
## [25] "PctBlack"             "PctAsian"             "PctOtherRace"
## [28] "PctMarriedHouseholds" "BirthRate"            "deathRate"
```

```
# Show summary
summary(cancer)
```

```
##        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
##
```

```r
# Show summary
length(cancer)
```

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

```r
# Show how many observations there are in each variable
for(df in cancer) {
  cat(sprintf("Dataframe length: %d\n", length(df)))
}
```

```
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
## Dataframe length: 3047
```

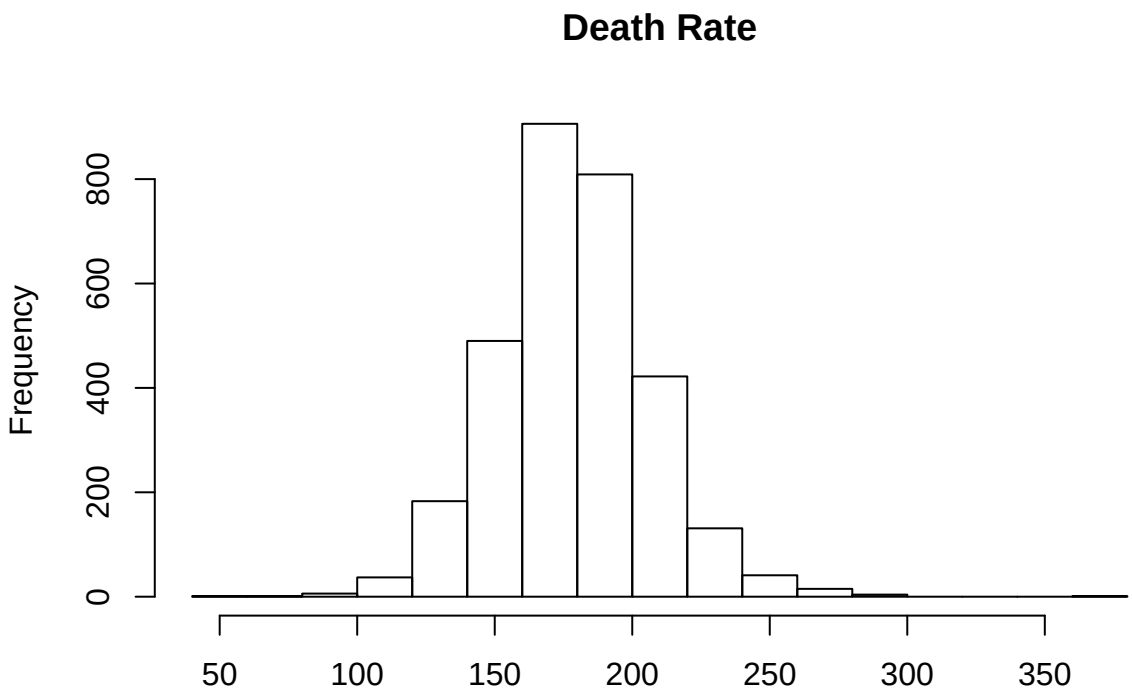
Univariate Analysis of Key Variables

```r
# Create a working copy of the variable cancer
canc <- cancer

# Summary
```

```r
summary(canc$deathRate)
```

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

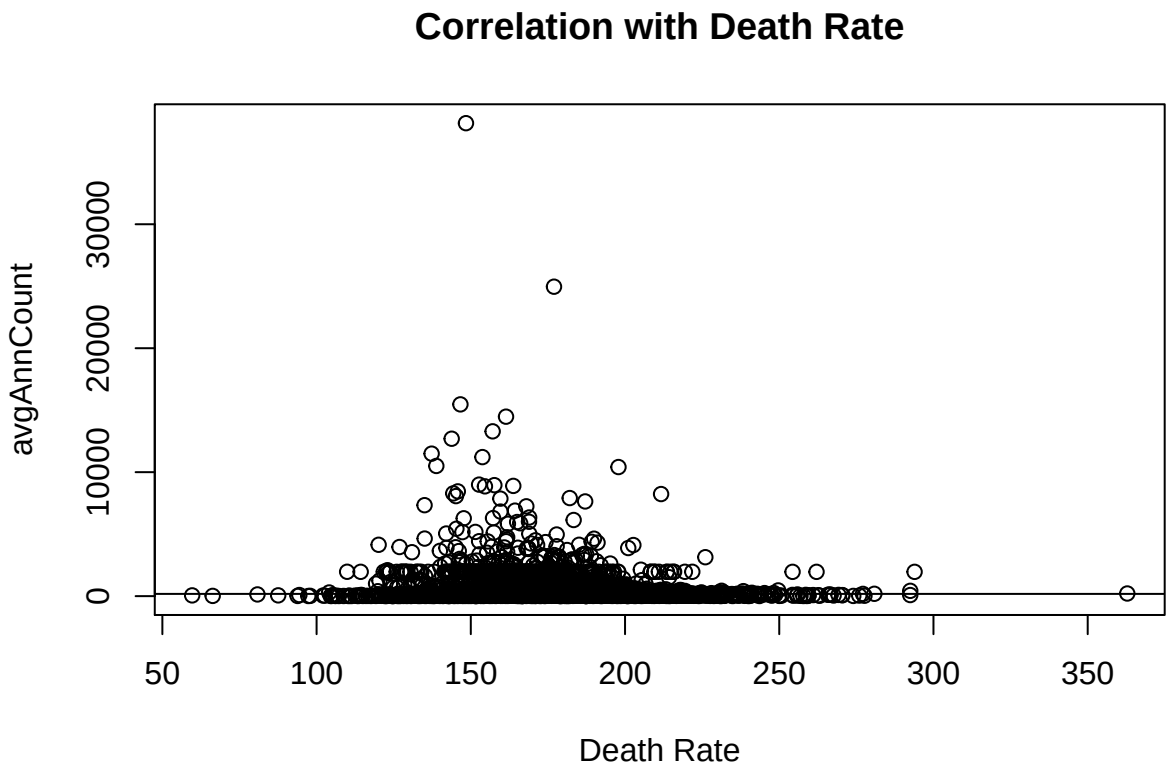
```r
# Histogram
hist(canc$deathRate, breaks = 20, main = "Death Rate", xlab = NULL)
```



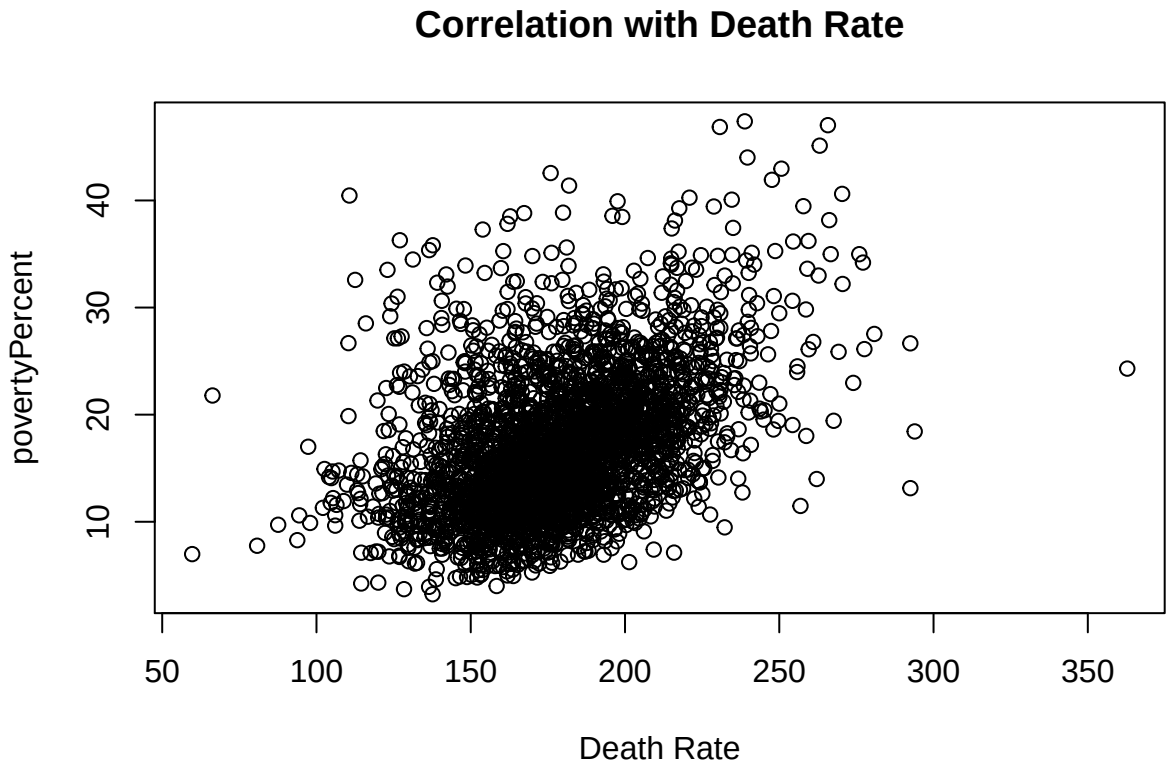

```r
# Define a function for running a set of compareative methods such as cor(), scattered plot, etc
compare_with_deathRate <- function(c, df, count) {
  if(is.na(colnames(c)[count])) {
    ylabel =  NULL
  } else {
    ylabel = colnames(c)[count]
  }

  # DEBUG
  #cat(sprintf("Dataframe name: %s\n", colnames(c)[df]))

  # Correlation
  cat(sprintf("Correlation - Death Rate and %s: %f\n", ylabel, cor(c$deathRate, df)))

  # Scatter Plot
  plot(jitter(c$deathRate, factor=2), jitter(df, factor=2),
       xlab = "Death Rate", ylab = ylabel,
       main = "Correlation with Death Rate")
  abline(lm(c$deathRate ~ df))
}

# Loop over all numeric variables in canc and analyze
count <- 1
cors = vector('numeric')
```

```
for(df in canc) {
  if(class(df) == "numeric" && colnames(canc)[count] != "deathRate") {
    compare_with_deathRate(canc, df, count)
    cors[count] = cor(canc$deathRate, df)
  }
  count = count + 1
}
```

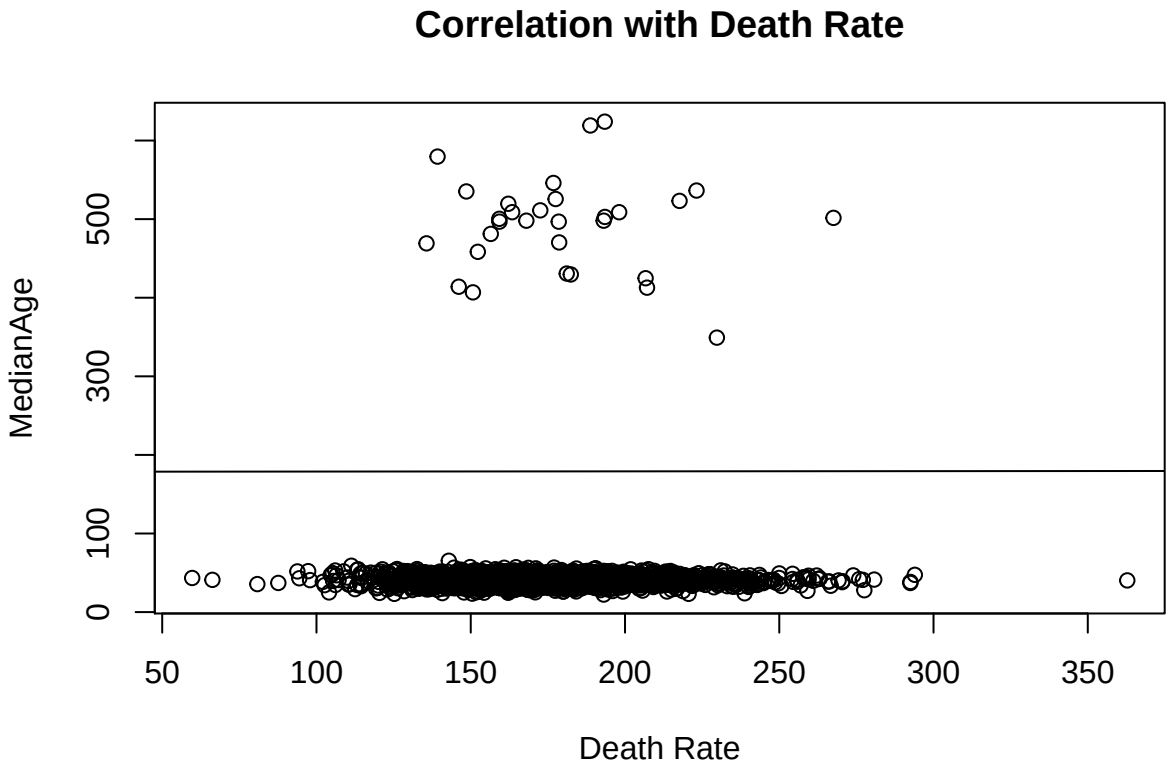
```
## Correlation - Death Rate and avgAnnCount: -0.143532
```



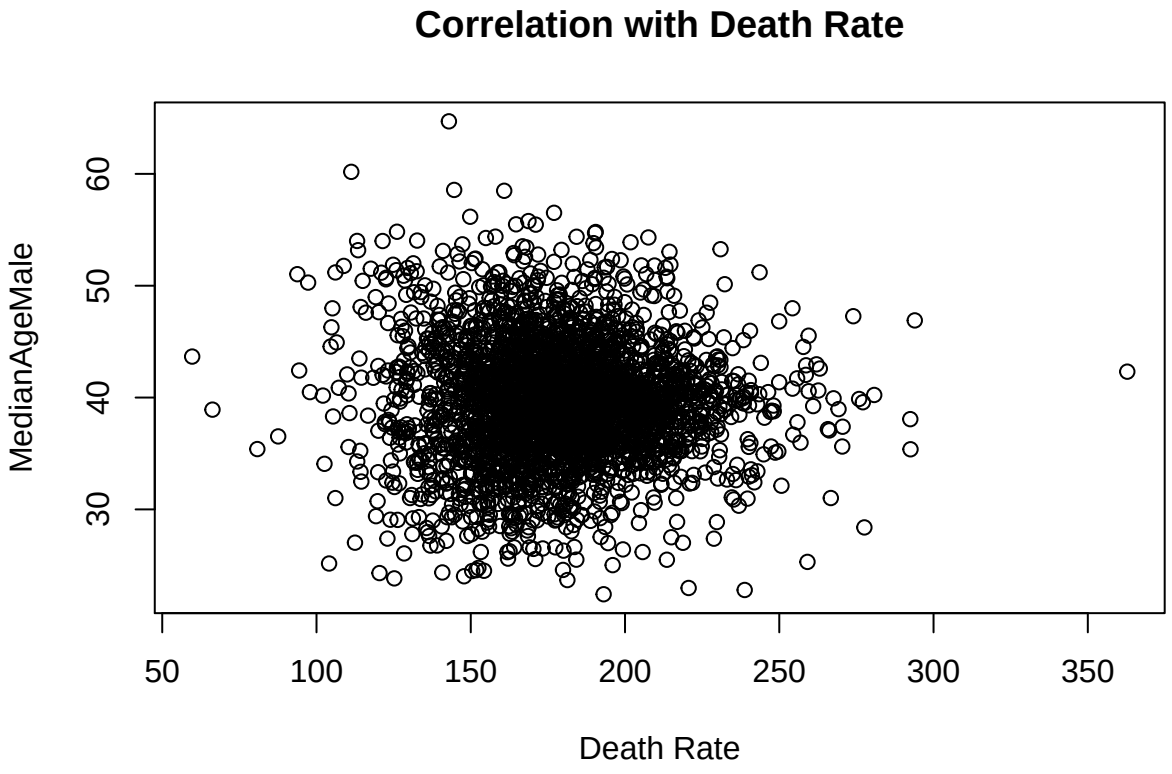

```
## Correlation - Death Rate and povertyPercent: 0.429389
```



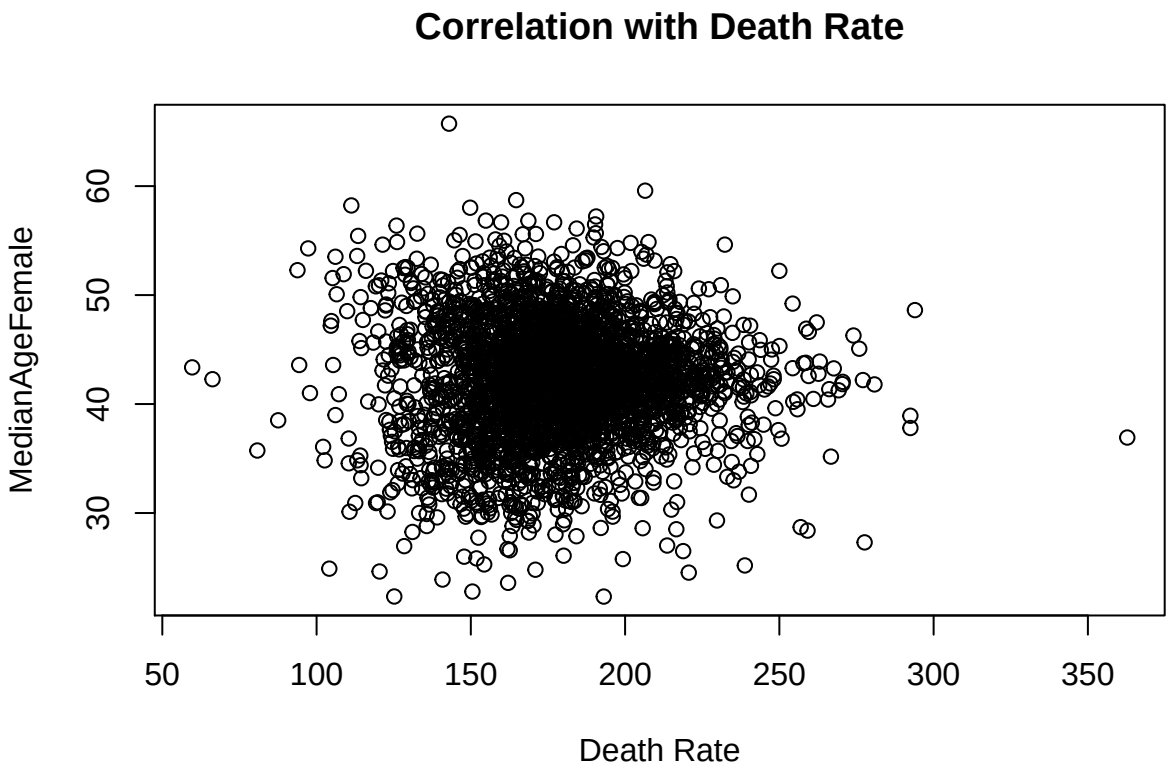

```
## Correlation - Death Rate and MedianAge: 0.004375
```



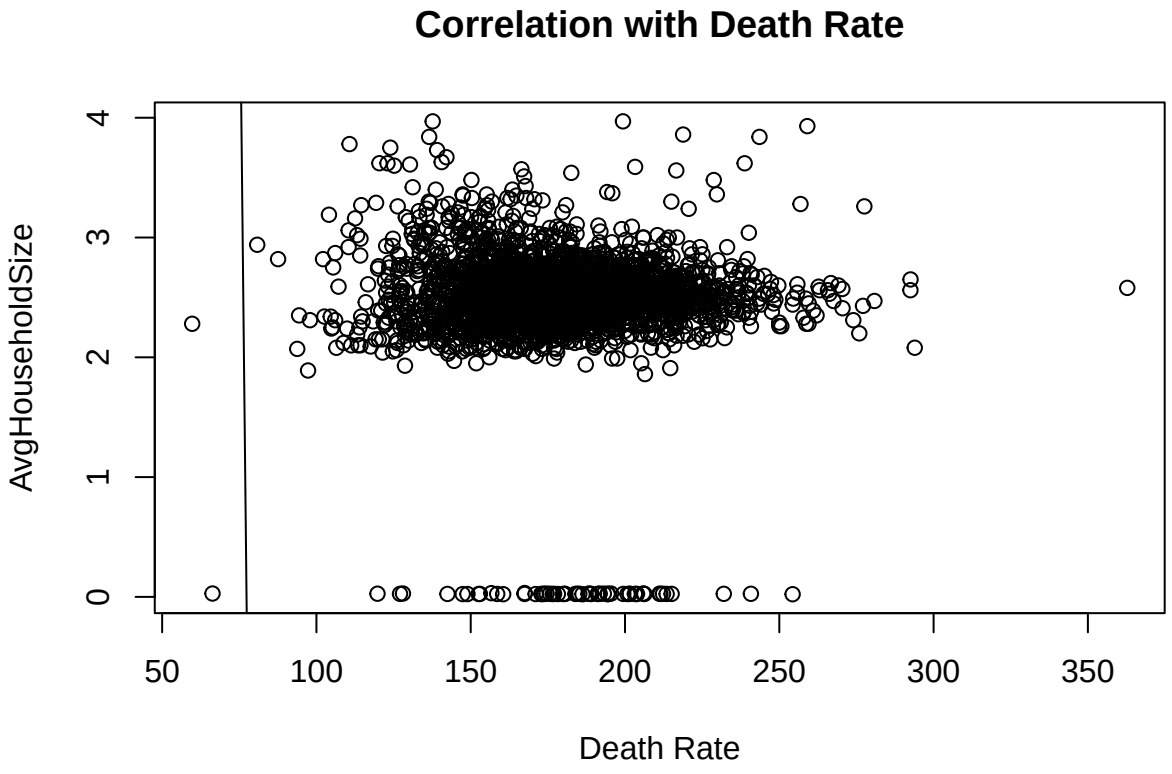

```
## Correlation - Death Rate and MedianAgeMale: -0.021929
```



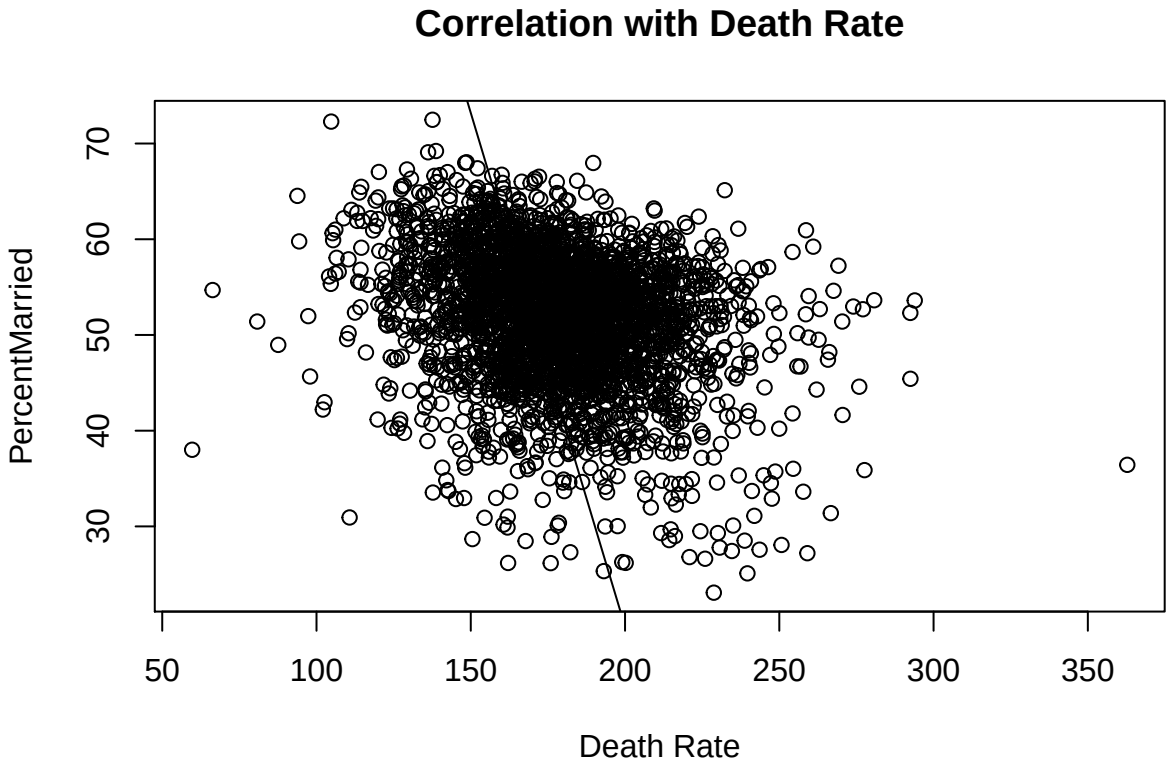

```
## Correlation - Death Rate and MedianAgeFemale: 0.012048
```



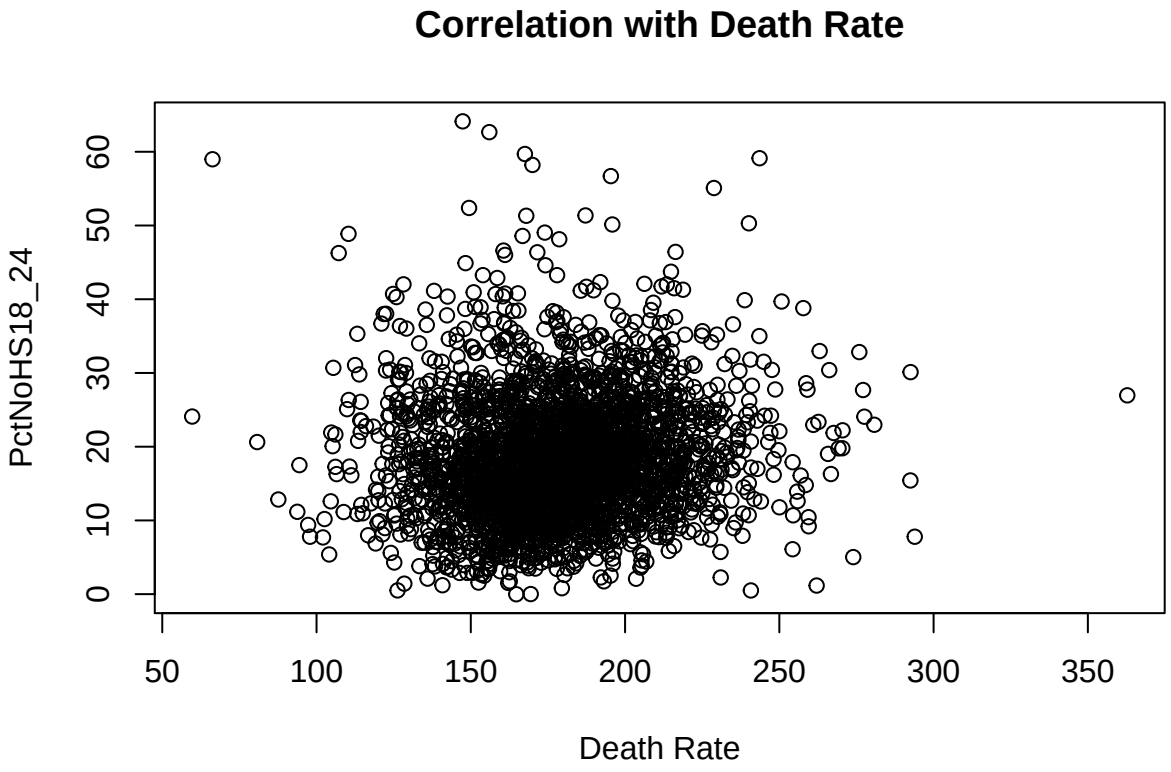

```
## Correlation - Death Rate and AvgHouseholdSize: -0.036905
```



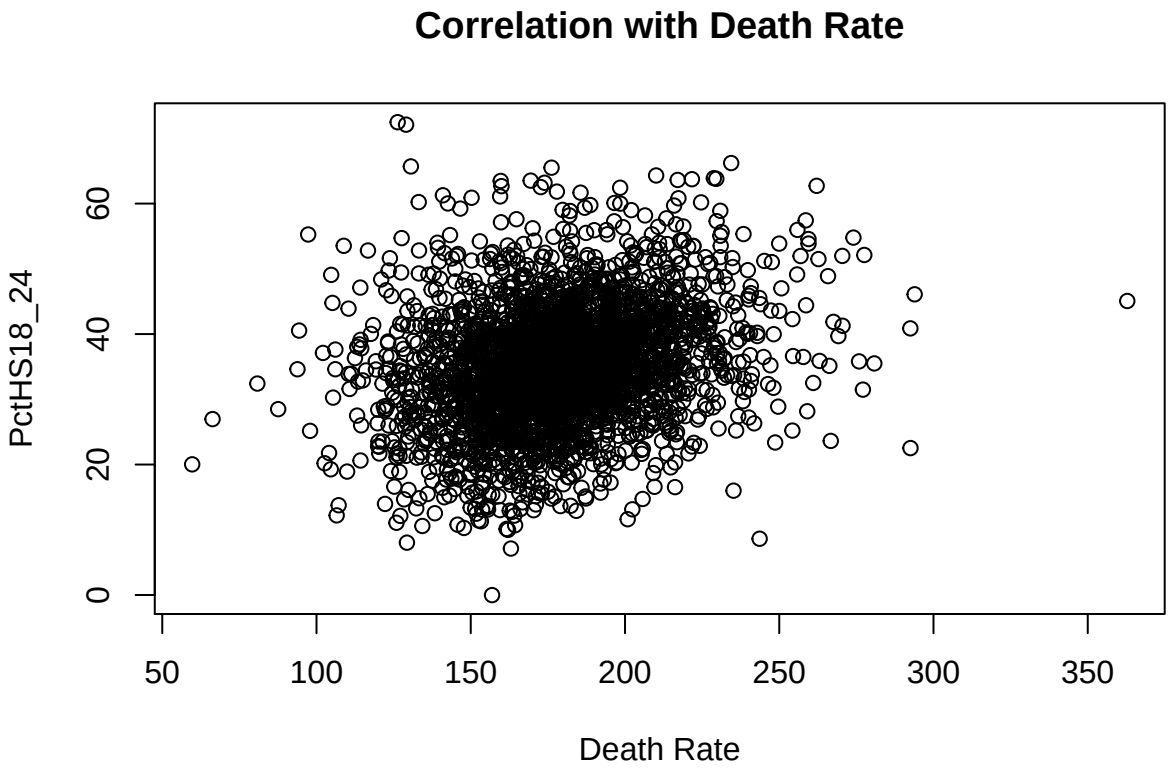

```
## Correlation - Death Rate and PercentMarried: -0.266820
```



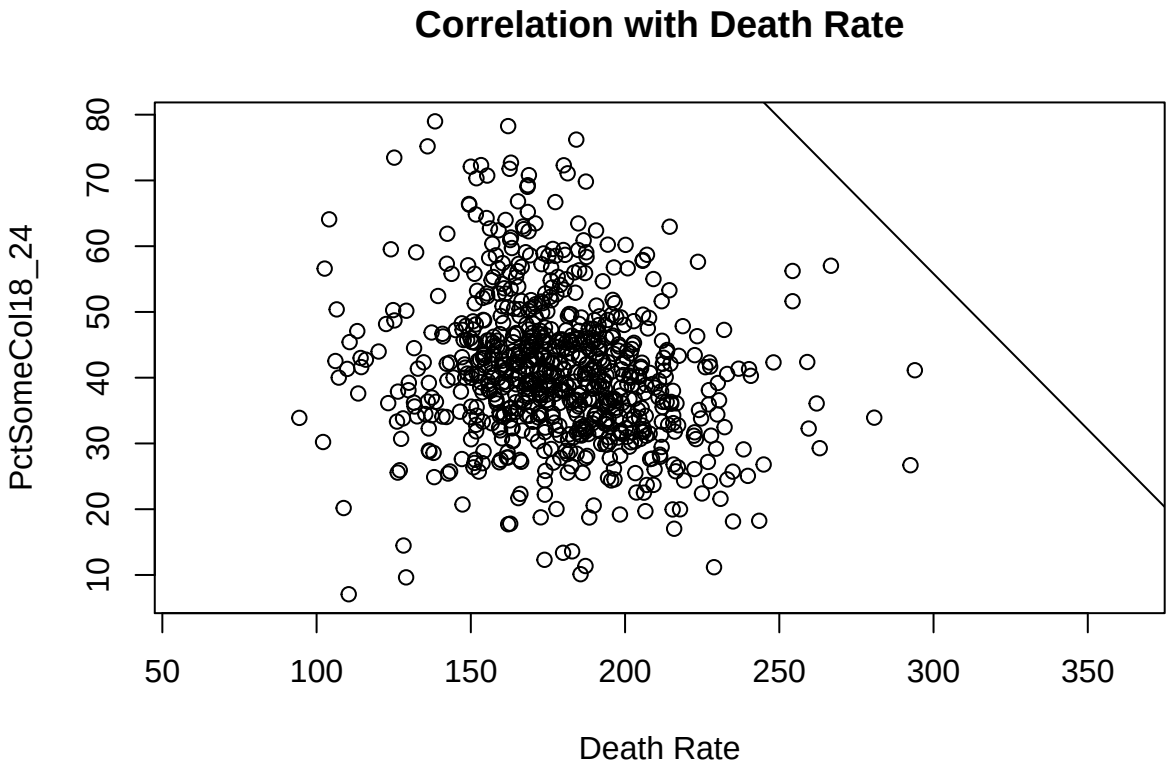

```
## Correlation - Death Rate and PctNoHS18_24: 0.088463
```



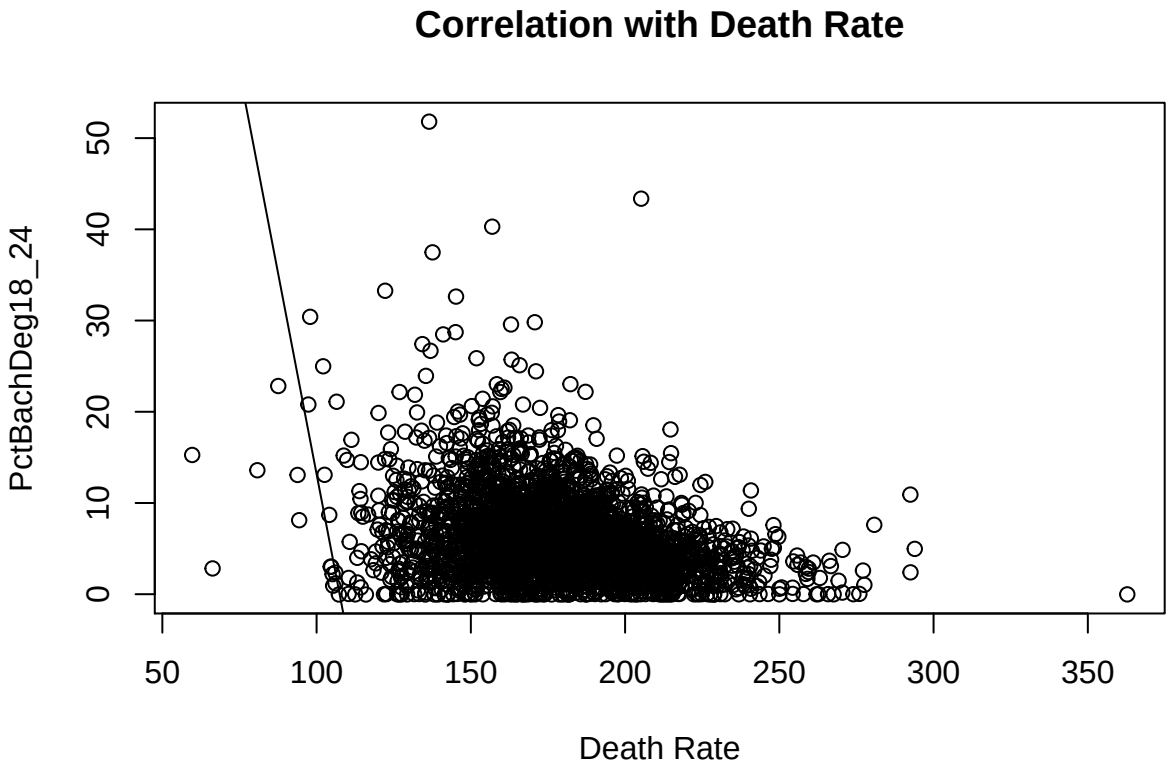

```
## Correlation - Death Rate and PctHS18_24: 0.261976
```



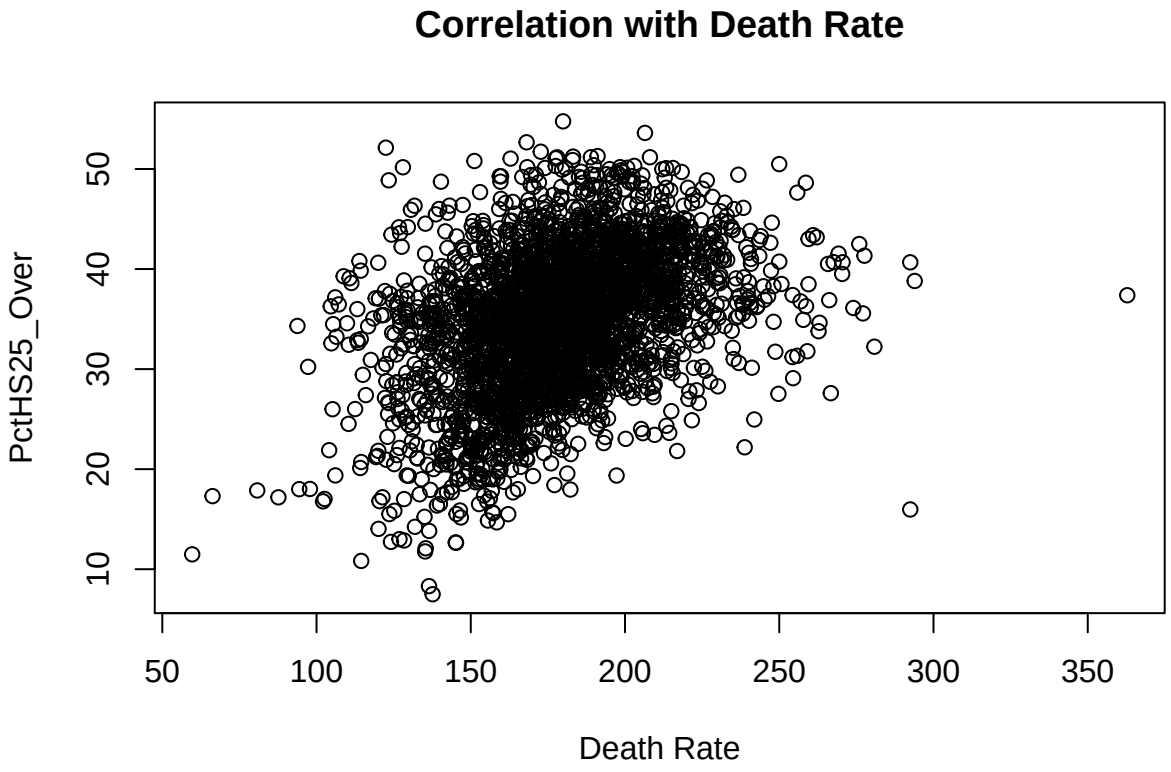

```
## Correlation - Death Rate and PctSomeCol18_24: NA
```



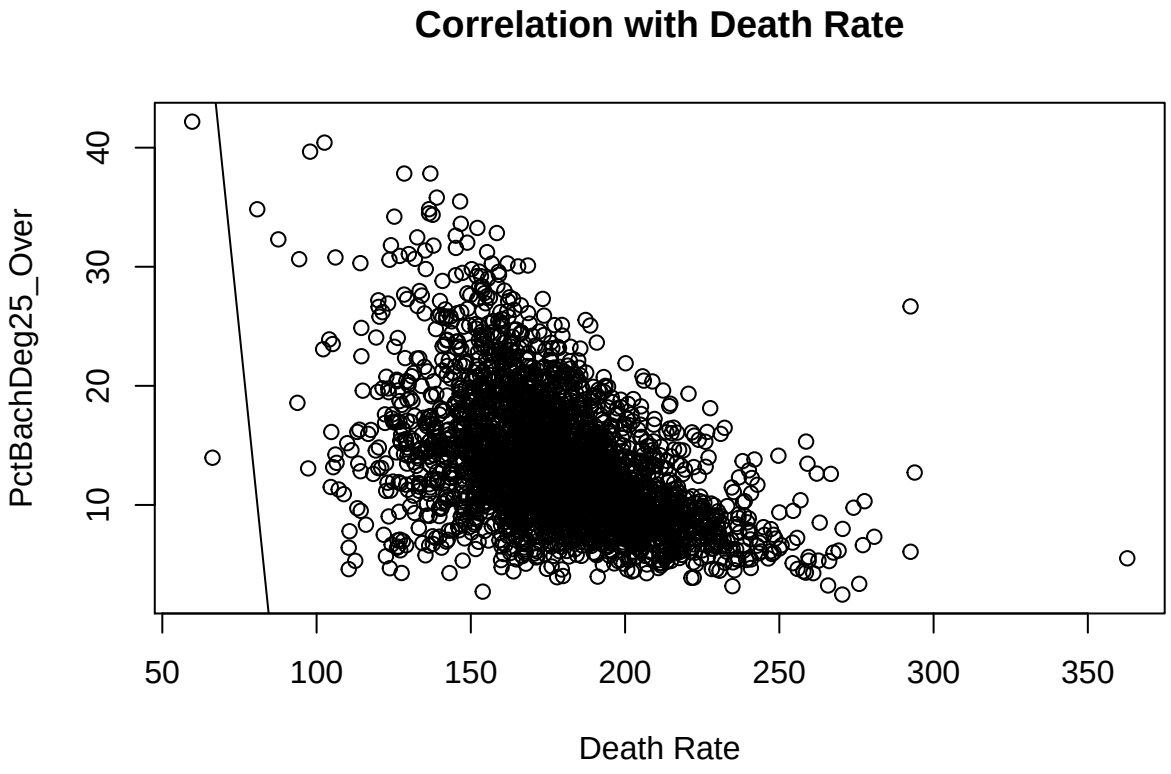

```
## Correlation - Death Rate and PctBachDeg18_24: -0.287817
```



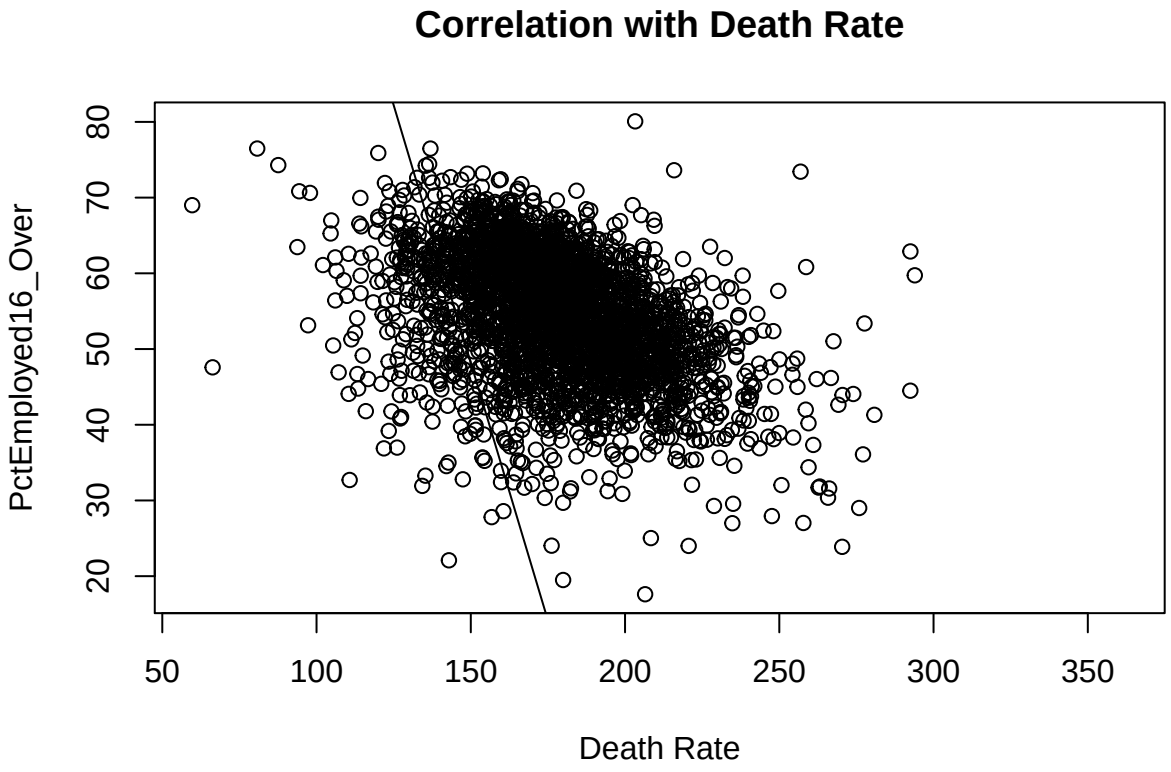

```
## Correlation - Death Rate and PctHS25_Over: 0.404589
```



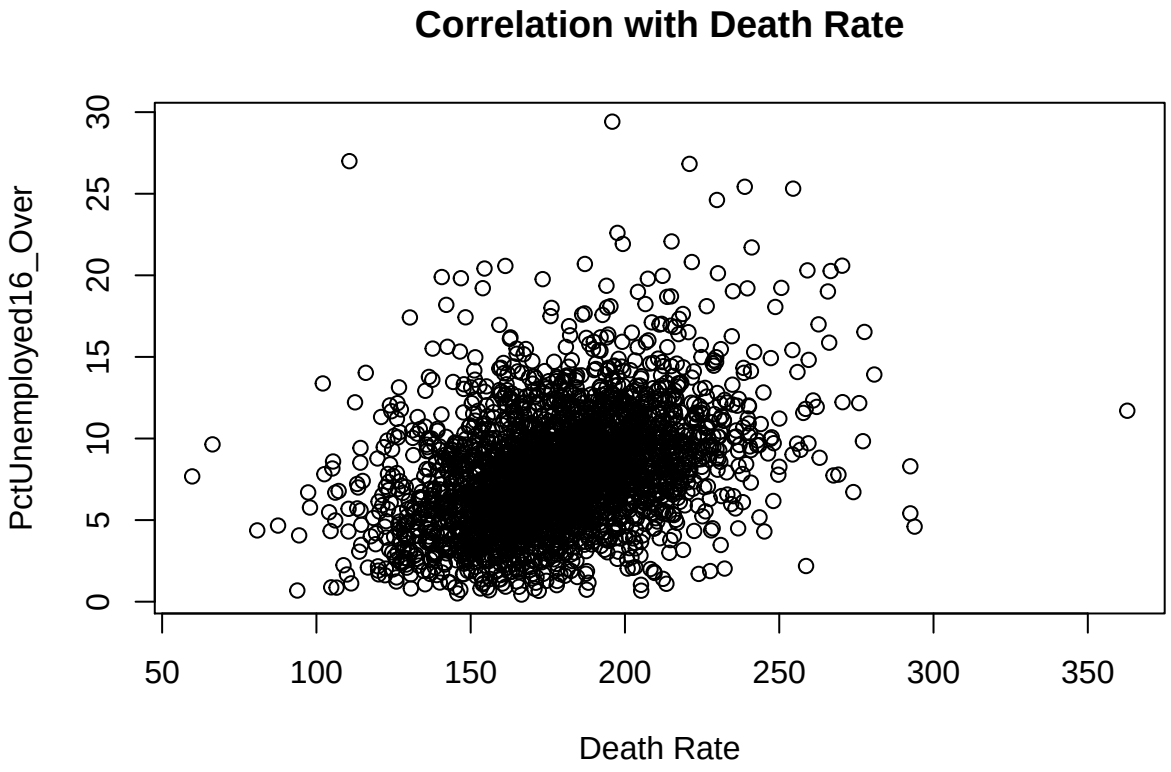

```
## Correlation - Death Rate and PctBachDeg25_Over: -0.485477
```



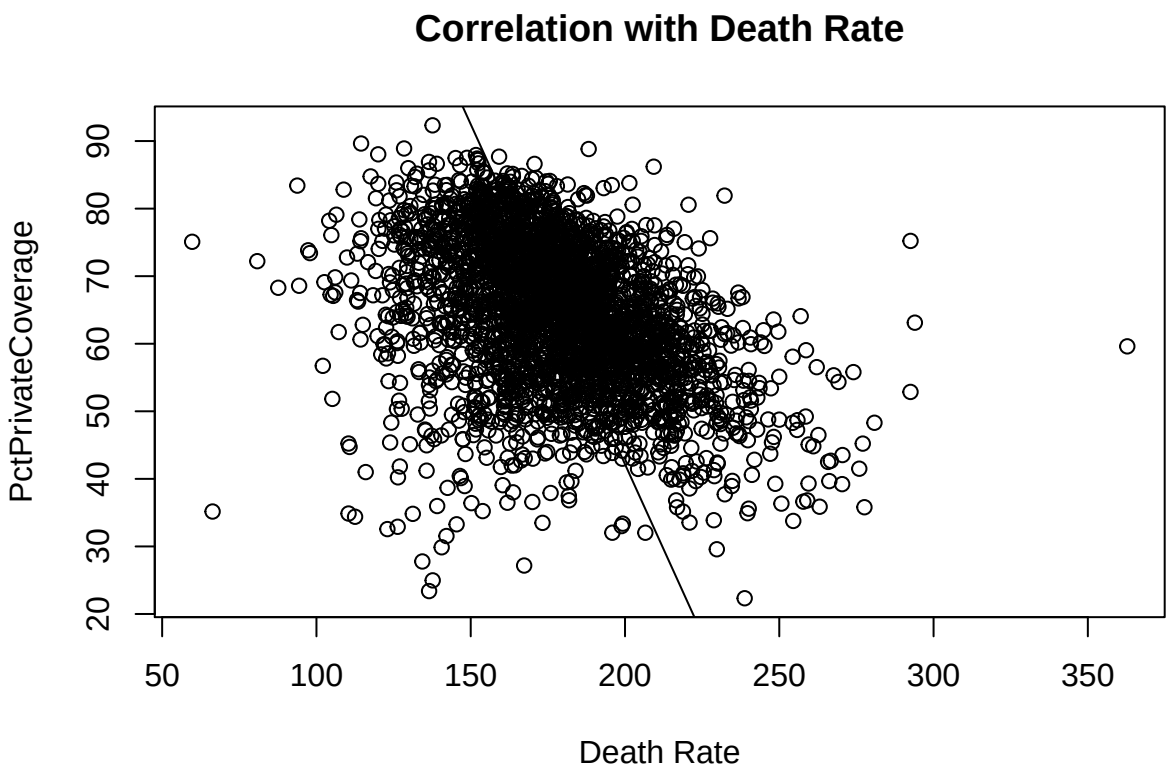

```
## Correlation - Death Rate and PctEmployed16_Over: NA
```



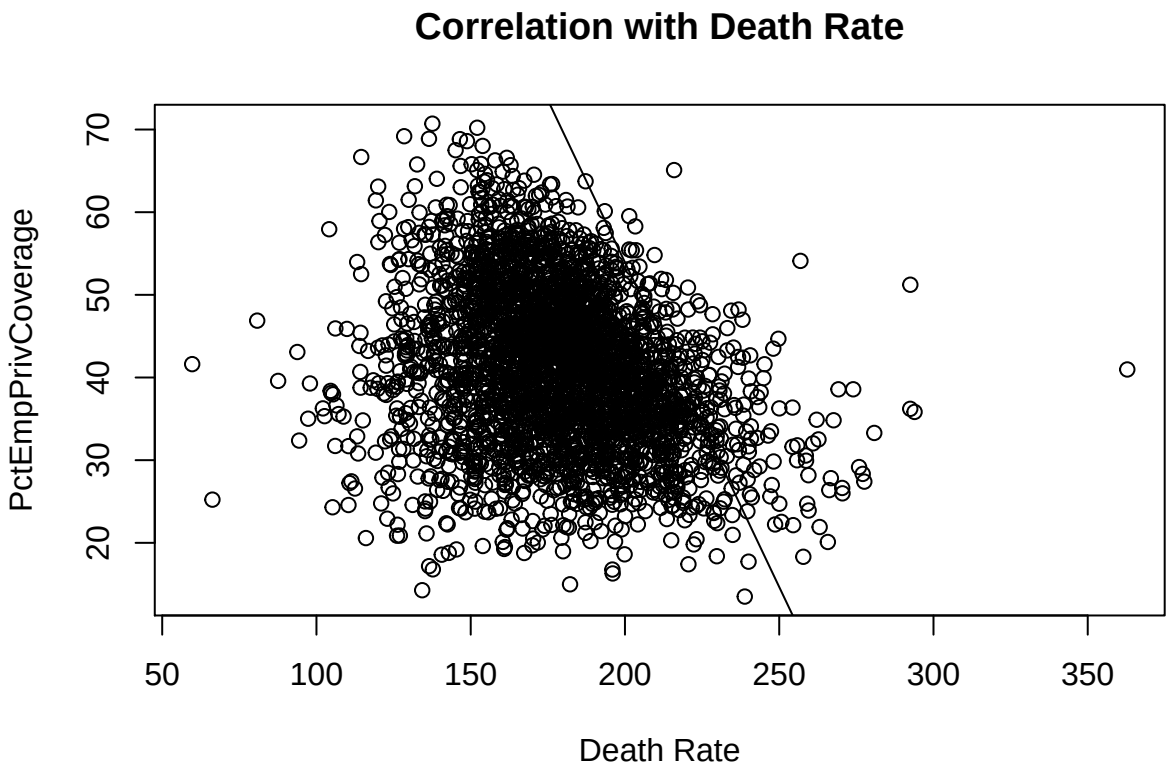

```
## Correlation - Death Rate and PctUnemployed16_Over: 0.378412
```



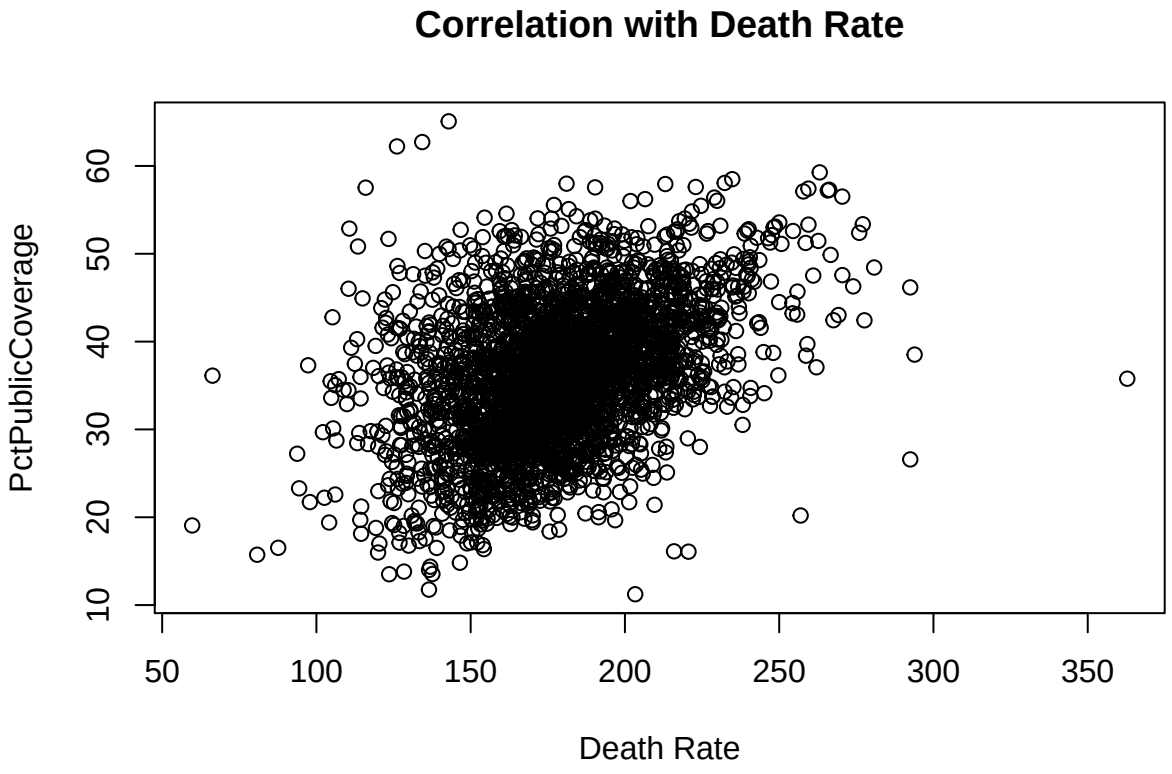

```
## Correlation - Death Rate and PctPrivateCoverage: -0.386066
```



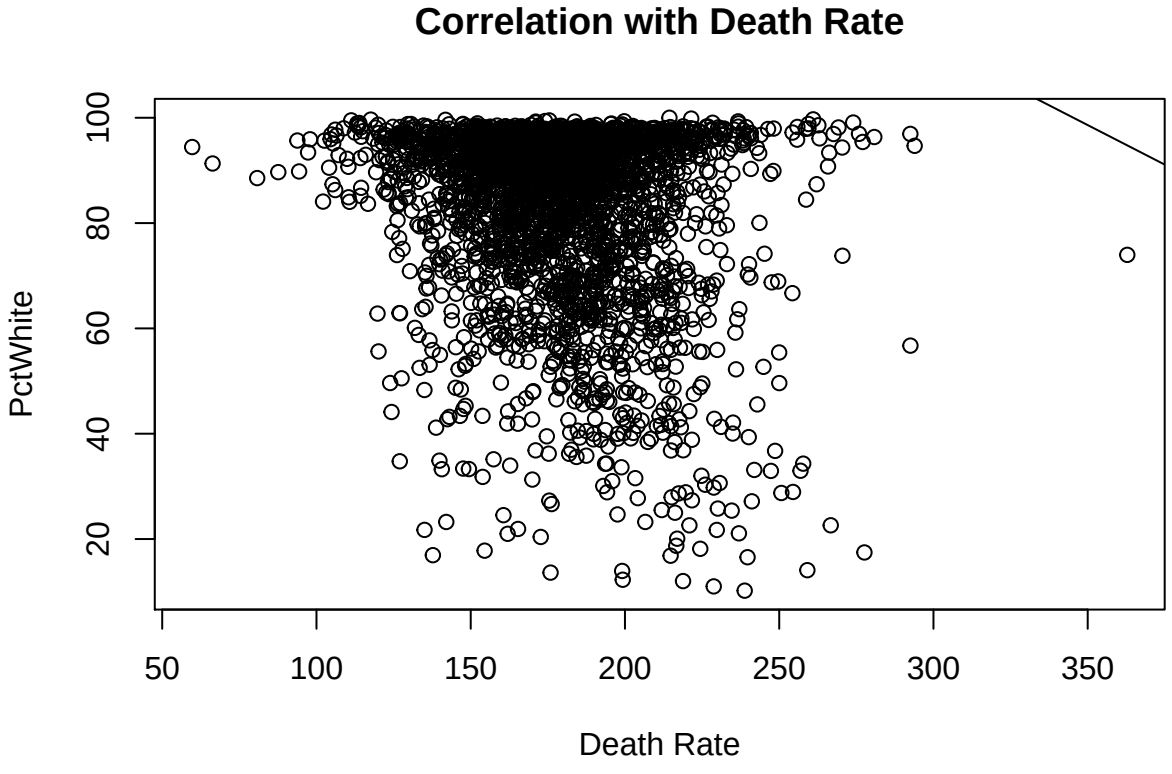

```
## Correlation - Death Rate and PctEmpPrivCoverage: -0.267399
```



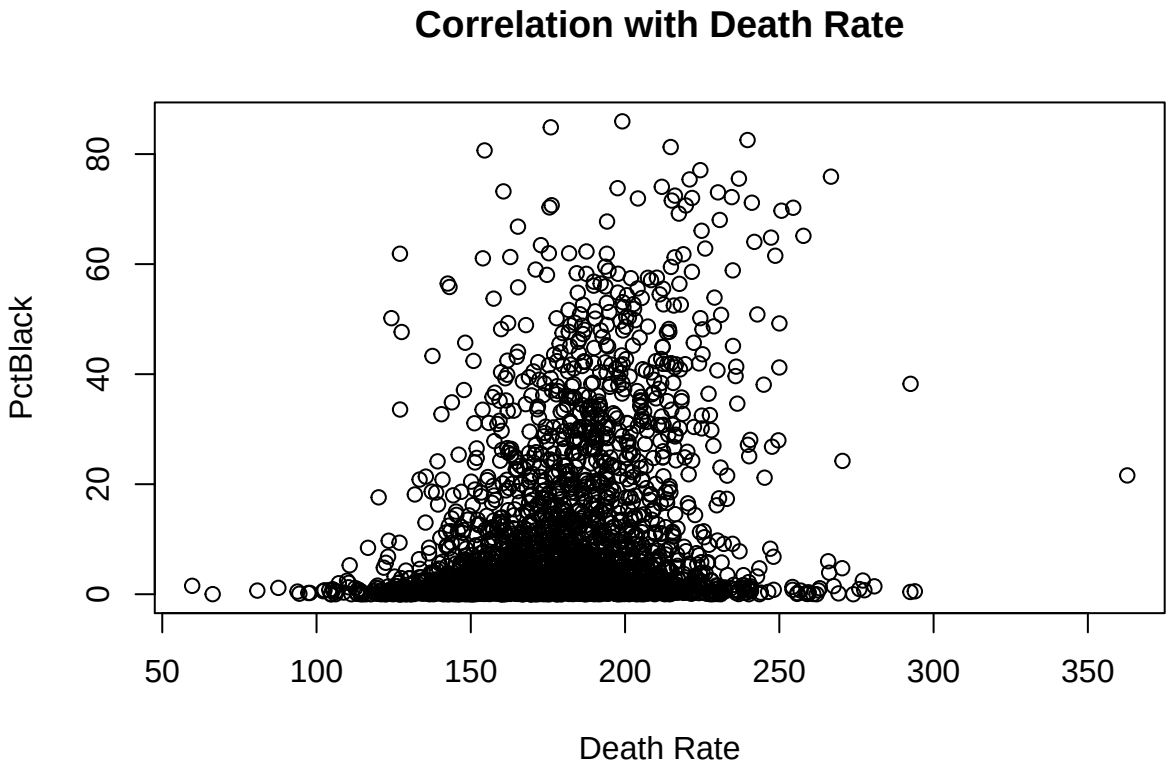

```
## Correlation - Death Rate and PctPublicCoverage: 0.404572
```



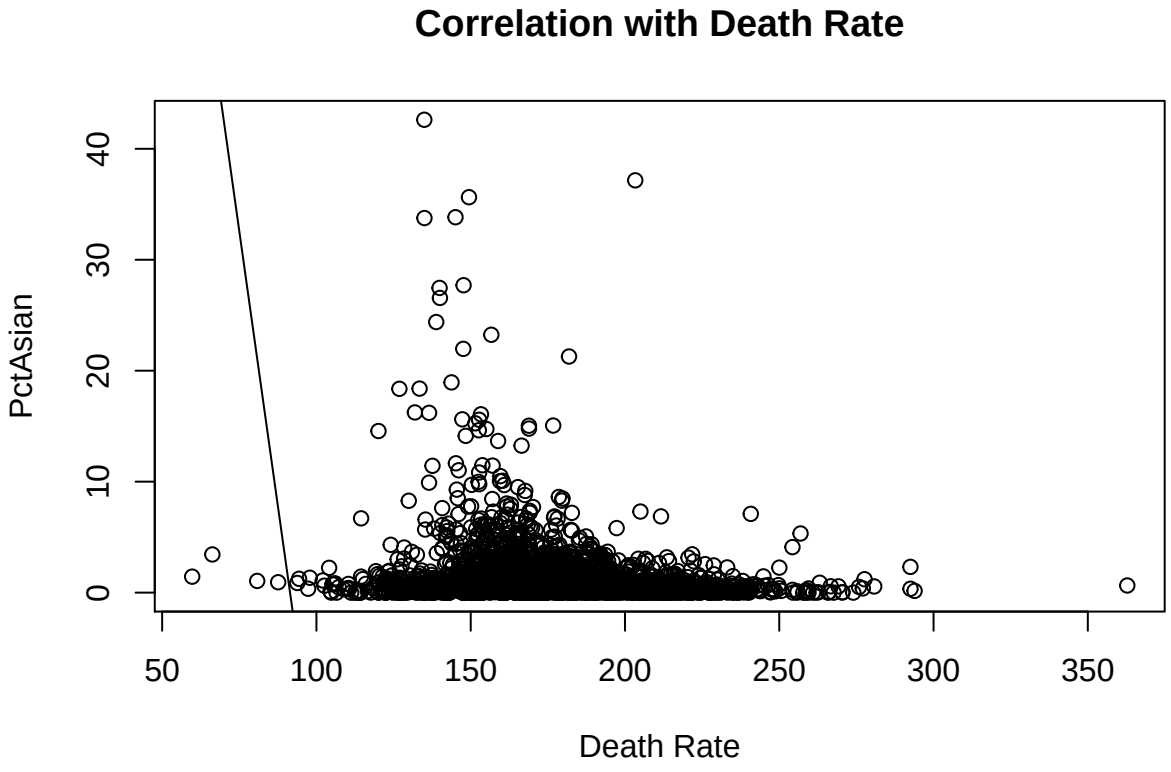

```
## Correlation - Death Rate and PctWhite: -0.177400
```



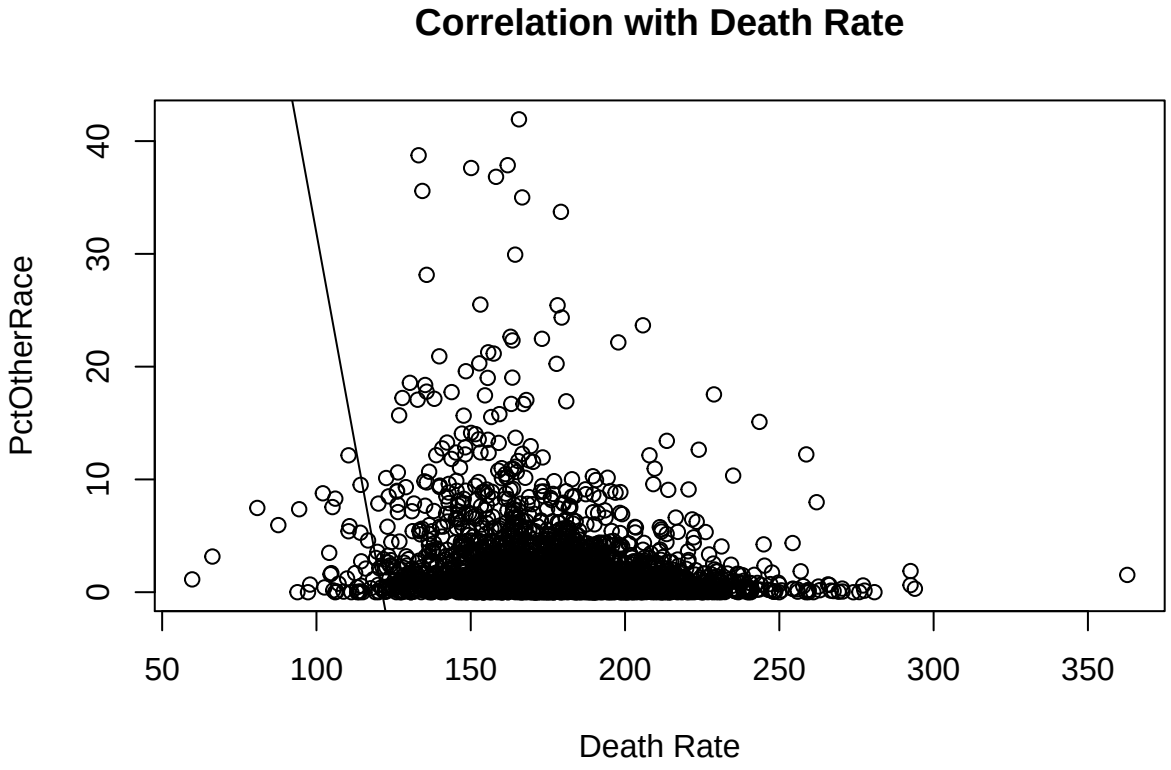

```
## Correlation - Death Rate and PctBlack: 0.257024
```



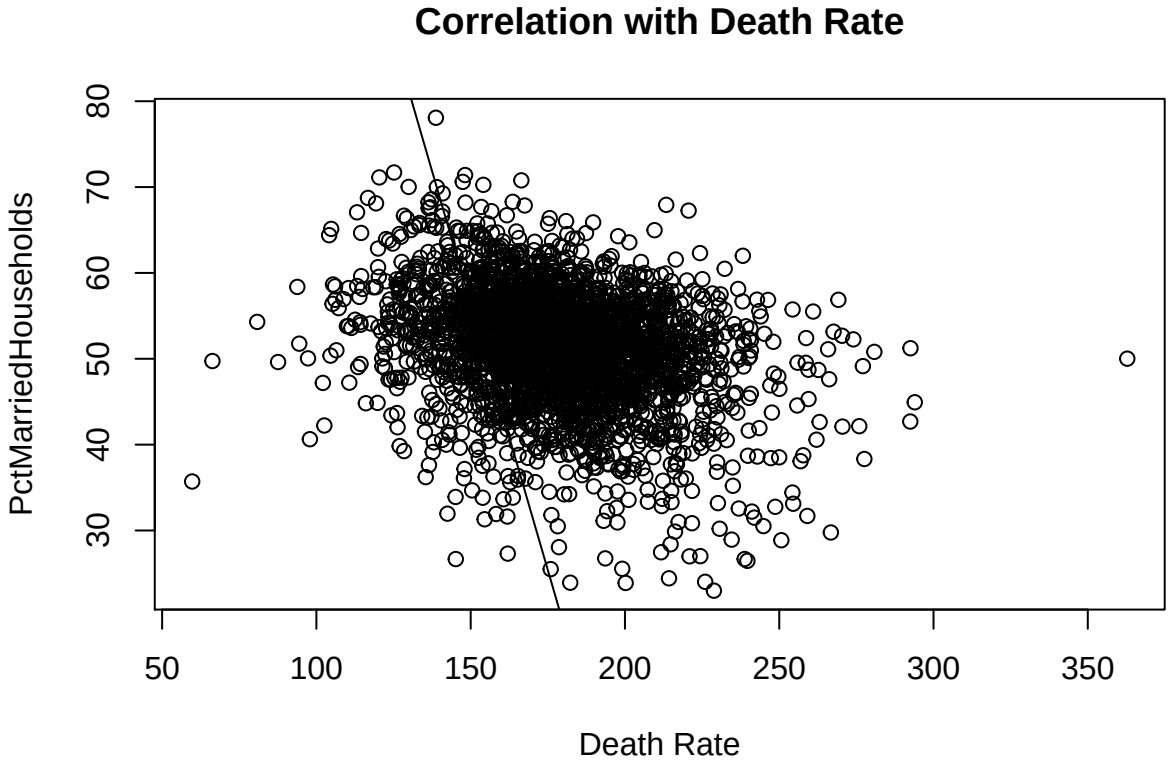

```
## Correlation - Death Rate and PctAsian: -0.186331
```



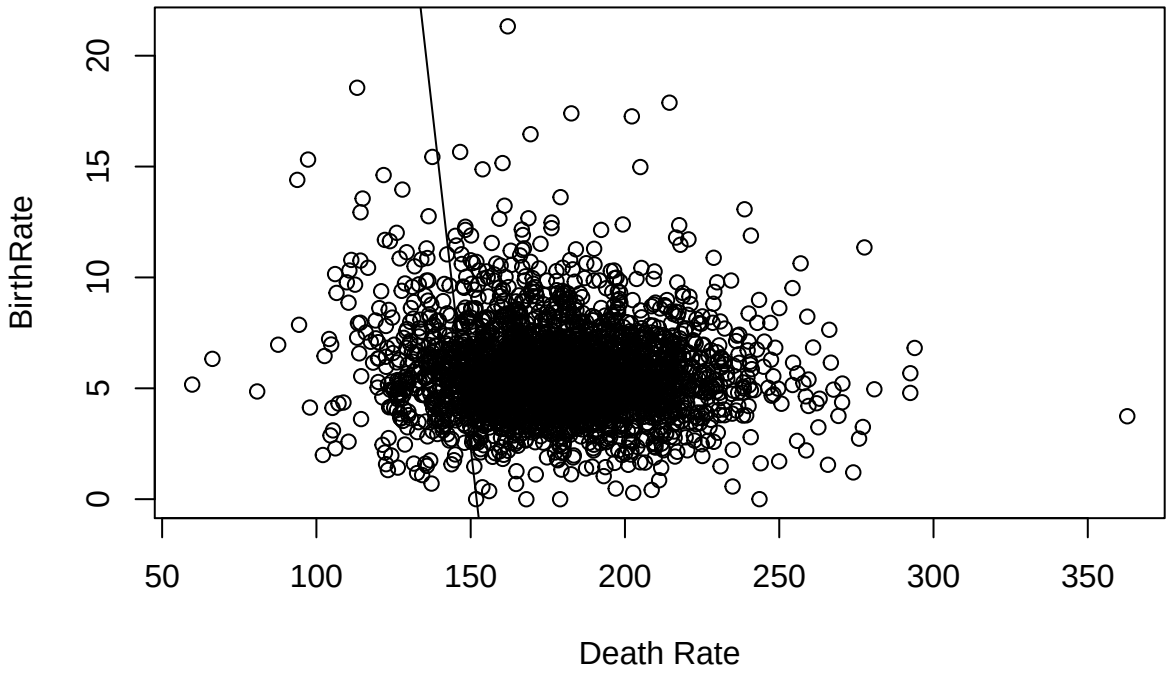

```
## Correlation - Death Rate and PctOtherRace: -0.189894
```




```
## Correlation - Death Rate and PctMarriedHouseholds: -0.293325
```




```
## Correlation - Death Rate and BirthRate: -0.087407
```




```
tail(sort(cors),5)
```

```
## [1] 0.2619759 0.3784124 0.4045717 0.4045891 0.4293890
```

```
head(sort(cors),5)
```

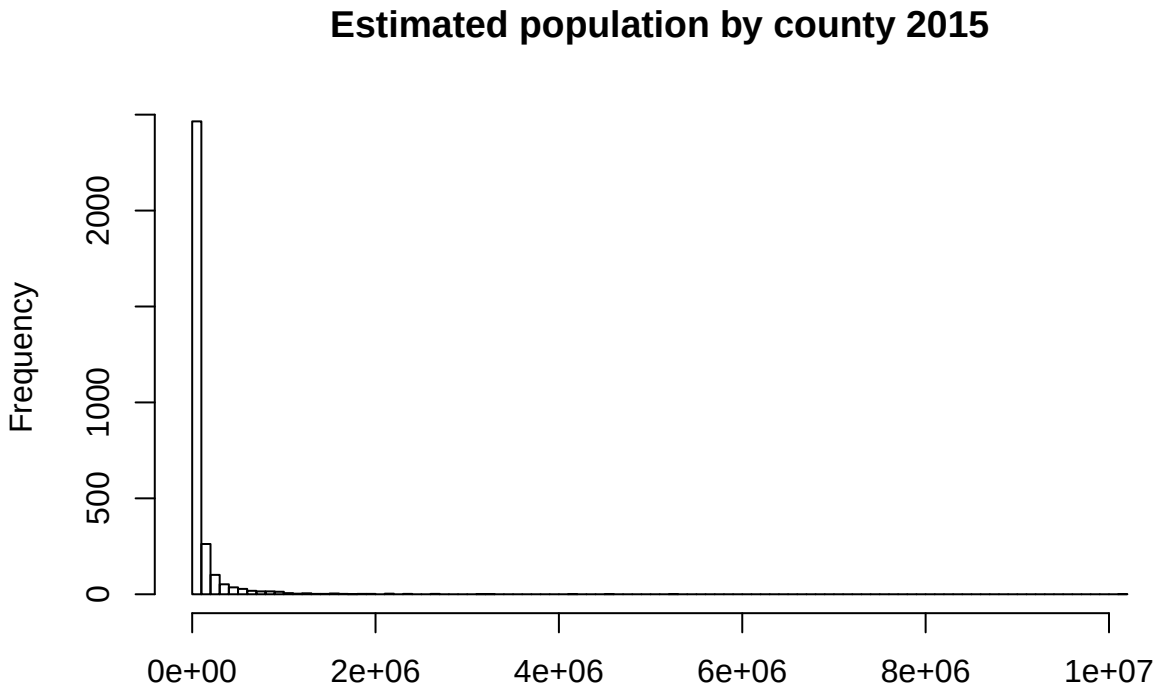
```
## [1] -0.4854773 -0.3860655 -0.2933253 -0.2878174 -0.2673994
```

Potentially erroneous data

```
# Estimated population by county 2015
summary(cancer$popEst2015)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      827    11684    26643   102637    68671 10170292
```

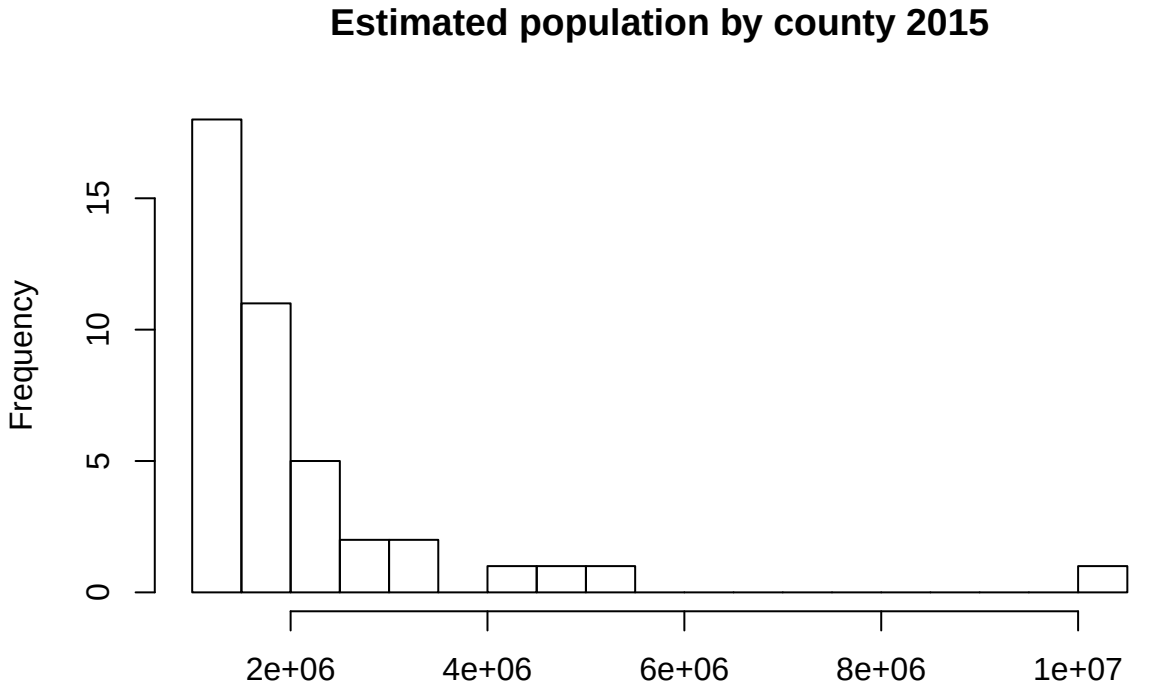
```
hist(cancer$popEst2015, breaks = 100, main = "Estimated population by county 2015", xlab = NULL)
```




```
# There are 42 countries with population over 1 million
over1m = subset(cancer, popEst2015 >= 1000000)
nrow(over1m)
```

```
## [1] 42
```

```
hist(over1m$popEst2015, breaks = 20, main = "Estimated population by county 2015", xlab = NULL)
```



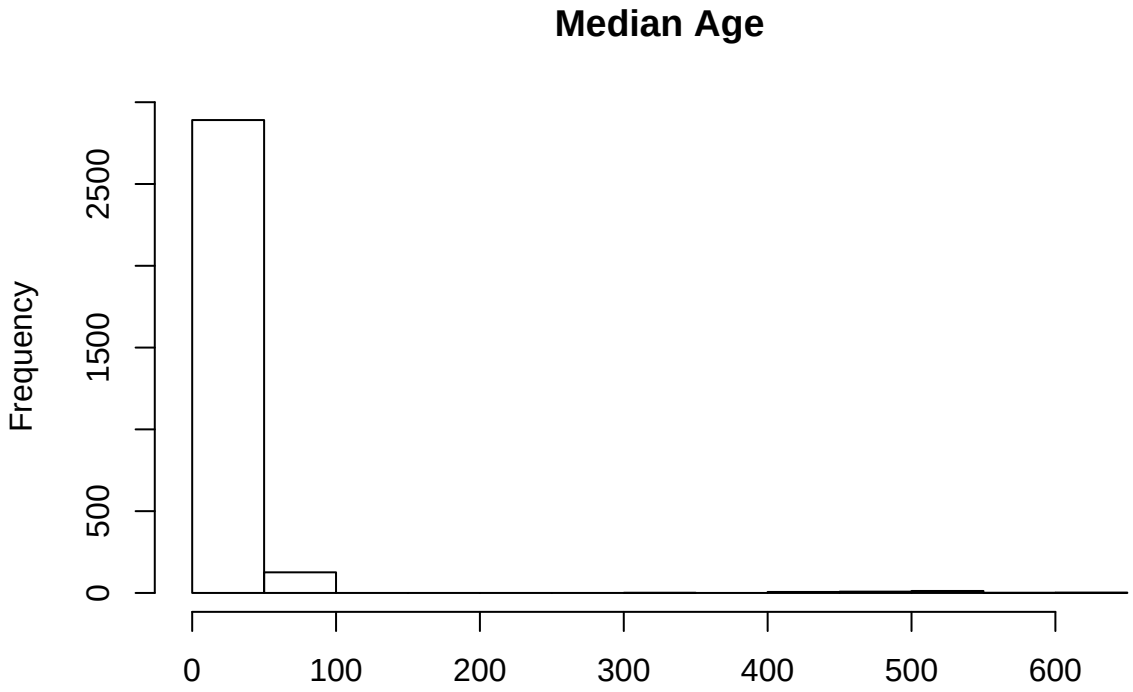

```
# One county has over 10 million
over10m = subset(cancer, popEst2015 >= 10000000)
nrow(over10m)
```

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

```
# Median Age
summary(cancer$MedianAge)
```

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

```
hist(cancer$MedianAge, breaks = 10, main = "Median Age", xlab = NULL)
```




```
# There are 30 countries with the median age over 100
medianAgeOver100 = subset(cancer, MedianAge >= 100)
nrow(medianAgeOver100)
```

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

Analysis of Key Relationships

```
# Focus on the variables that exihibited stronger relationships with deathRate

## Top 3 Positive Correlations

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

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

```
cor(cancer$deathRate, cancer$PctHS25_Over)
```

```
## [1] 0.4045891
```

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

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

```r
## Top 3 Negative Correlations
cor(cancer$deathRate, cancer$PctBachDeg25_Over)
```

```
## [1] -0.4854773
```

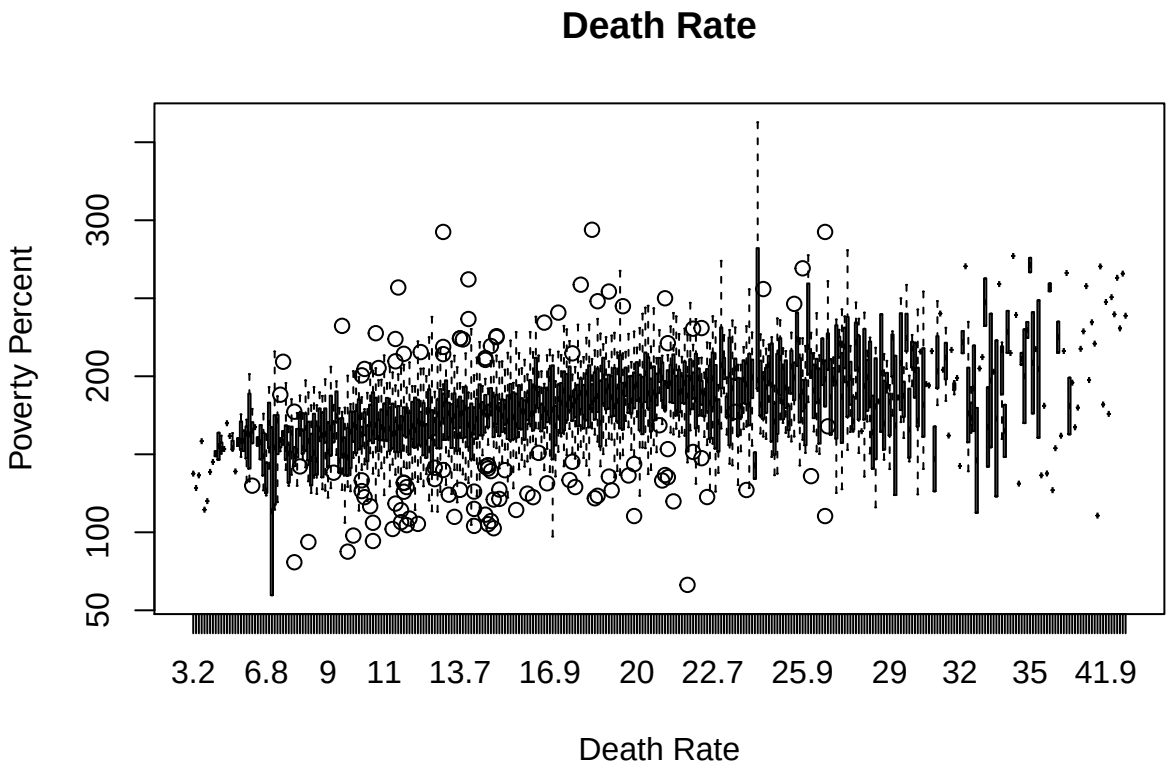
```r
cor(cancer$deathRate, cancer$PctPrivateCoverage)
```

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

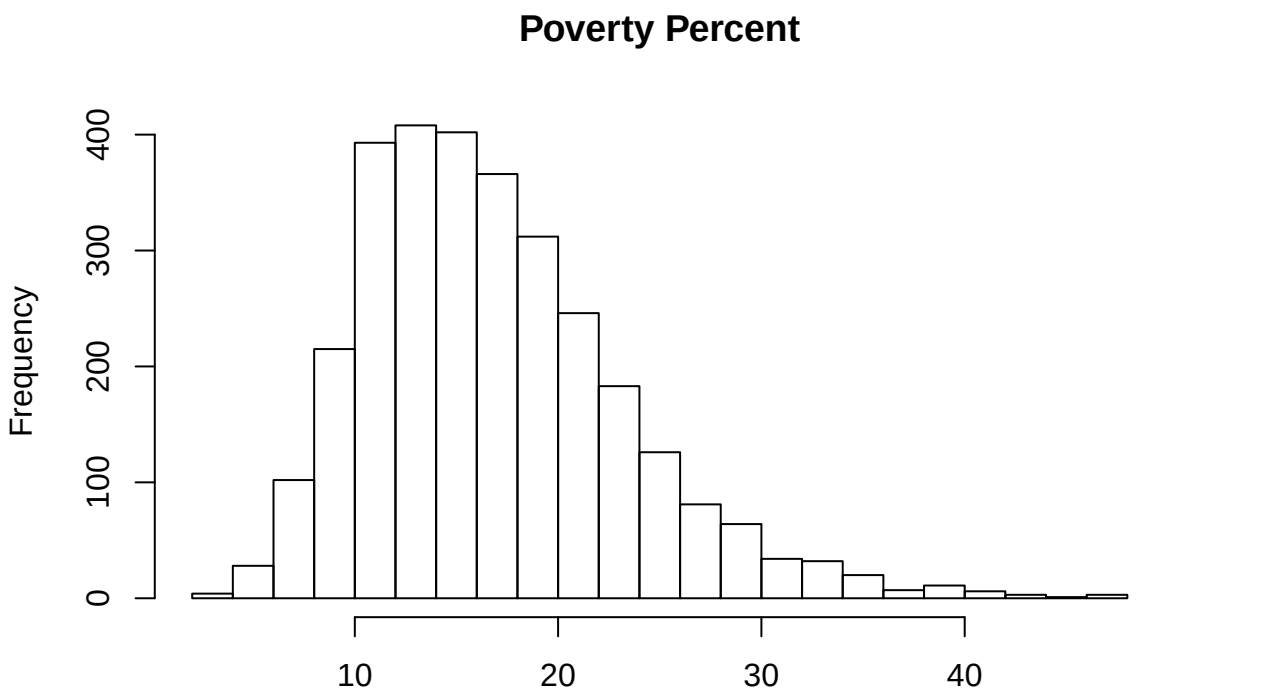
```r
cor(cancer$deathRate, cancer$PctMarriedHouseholds)
```

```
## [1] -0.2933253
```

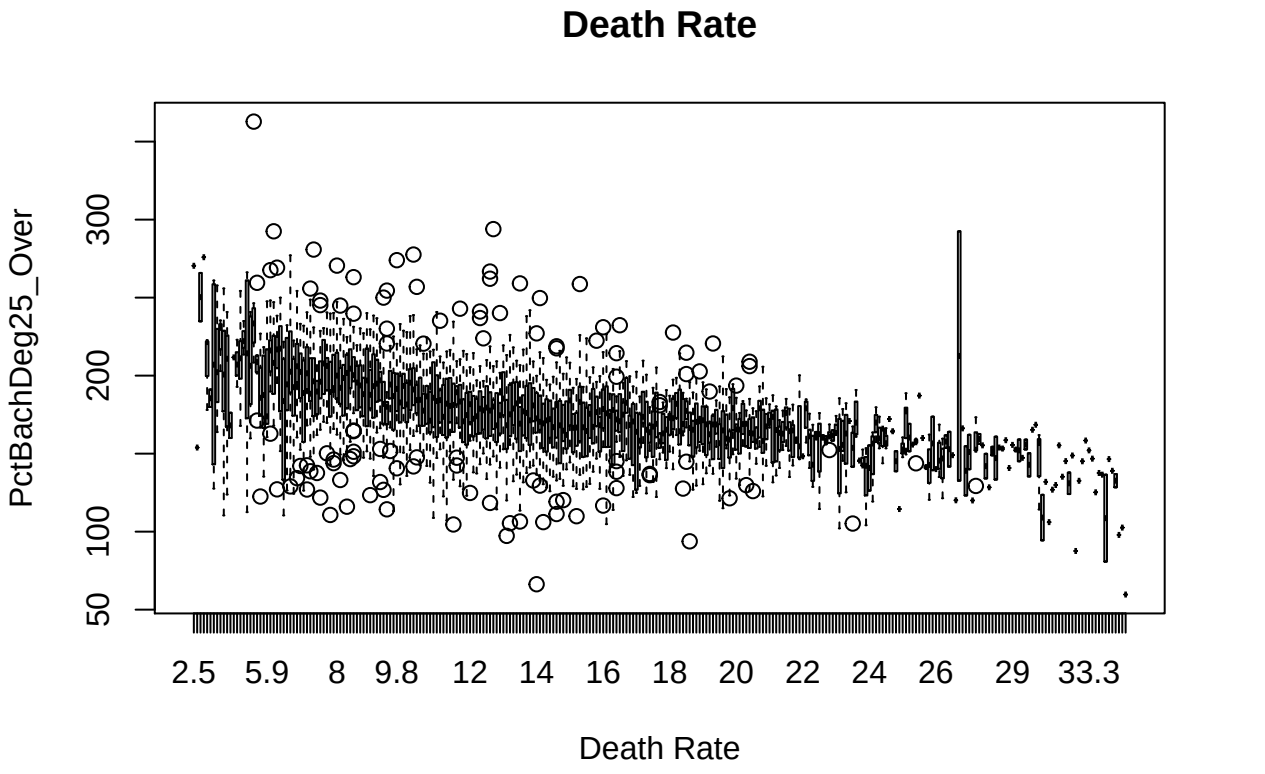
```r
# Box Plots
boxplot(deathRate ~ povertyPercent, data = canc,
        main = "Death Rate",
  xlab = "Death Rate", ylab = "Poverty Percent")
```



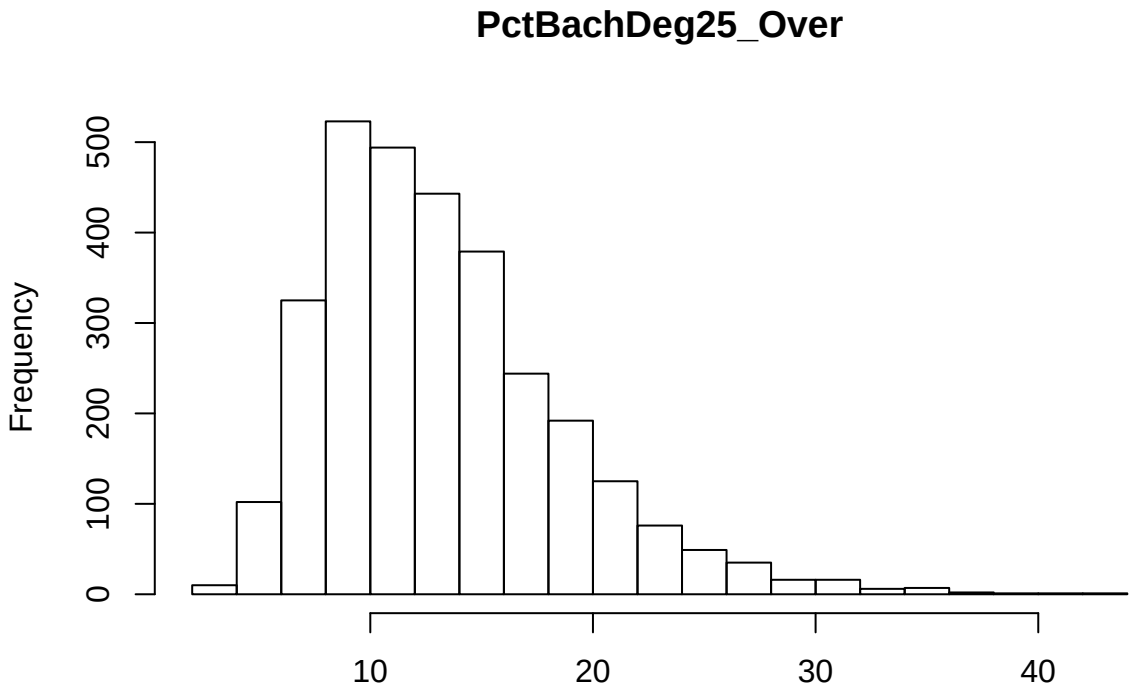

```r
hist(cancer$povertyPercent, breaks = 20, main = "Poverty Percent", xlab = NULL)
```




```
boxplot(deathRate ~ PctBachDeg25_Over, data = canc,
        main = "Death Rate",
  xlab = "Death Rate", ylab = "PctBachDeg25_Over")
```




```
hist(cancer$PctBachDeg25_Over, breaks = 20, main = "PctBachDeg25_Over", xlab = NULL)
```



Objective

Perform an exploratory analysis to understand how county-level characteristics are related to cancer mortality.