An Exploratory Analysis of Cancer Incidence and Mortality to Identify High-Risk Communities and Improve Survival

Ramiro Cadavid, Pri Nonis, Payman Roghani

September 24, 2018

Setup

Introduction

In this project our efforts are focused on the analysis of data included in the csv file provided, to primarily understand the potential relationship between different parameters and the incidences of cancer across counties in the US. The main objectives are: >1. To understand factors that predict cancer mortality rate, with the ultimate aim of identifying communities for social interventions. >2. To determine which interventions are likely to have the most impact.

Cancer Data

```
Cancer <- read.csv('cancer.csv', row.names = 1)
```

```
summary(Cancer) #summary statistics
  avgAnnCount        medIncome        popEst2015       povertyPercent
 Min.   :    6.0   Min.   : 22640   Min.   :     827   Min.   : 3.20
 1st Qu.:   76.0   1st Qu.: 38882   1st Qu.:   11684   1st Qu.:12.15
 Median :  171.0   Median : 45207   Median :   26643   Median :15.90
 Mean   :  606.3   Mean   : 47063   Mean   :  102637   Mean   :16.88
 3rd Qu.:  518.0   3rd Qu.: 52492   3rd Qu.:   68671   3rd Qu.:20.40
 Max.   :38150.0   Max.   :125635   Max.   :10170292   Max.   :47.40

              binnedInc      MedianAge      MedianAgeMale
 (45201, 48021.6]  : 306   Min.   : 22.30   Min.   :22.40
 (54545.6, 61494.5]: 306   1st Qu.: 37.70   1st Qu.:36.35
 [22640, 34218.1]  : 306   Median : 41.00   Median :39.60
 (42724.4, 45201]  : 305   Mean   : 45.27   Mean   :39.57
 (48021.6, 51046.4]: 305   3rd Qu.: 44.00   3rd Qu.:42.50
 (51046.4, 54545.6]: 305   Max.   :624.00   Max.   :64.70
 (Other)           :1214
 MedianAgeFemale                            Geography    AvgHouseholdSize
 Min.   :22.30   Abbeville County, South Carolina:   1   Min.   :0.0221
 1st Qu.:39.10   Acadia Parish, Louisiana        :   1   1st Qu.:2.3700
 Median :42.40   Accomack County, Virginia       :   1   Median :2.5000
 Mean   :42.15   Ada County, Idaho               :   1   Mean   :2.4797
 3rd Qu.:45.30   Adair County, Iowa              :   1   3rd Qu.:2.6300
 Max.   :65.70   Adair County, Kentucky          :   1   Max.   :3.9700
                 (Other)                         :3041
 PercentMarried   PctNoHS18_24     PctHS18_24   PctSomeCol18_24
 Min.   :23.10   Min.   : 0.00   Min.   : 0.0   Min.   : 7.10
 1st Qu.:47.75   1st Qu.:12.80   1st Qu.:29.2   1st Qu.:34.00  
 Median :52.40   Median :17.10   Median :34.7   Median :40.40  
 Mean   :51.77   Mean   :18.22   Mean   :35.0   Mean   :40.98  
 3rd Qu.:56.40   3rd Qu.:22.70   3rd Qu.:40.7   3rd Qu.:46.40  
 Max.   :72.50   Max.   :64.10   Max.   :72.5   Max.   :79.00  
                                                NA's   :2285   
 PctBachDeg18_24   PctHS25_Over   PctBachDeg25_Over PctEmployed16_Over
 Min.   : 0.000   Min.   : 7.50   Min.   : 2.50     Min.   :17.60     
 1st Qu.: 3.100   1st Qu.:30.40   1st Qu.: 9.40     1st Qu.:48.60     
 Median : 5.400   Median :35.30   Median :12.30     Median :54.50     
 Mean   : 6.158   Mean   :34.80   Mean   :13.28     Mean   :54.15     
 3rd Qu.: 8.200   3rd Qu.:39.65   3rd Qu.:16.10     3rd Qu.:60.30     
 Max.   :51.800   Max.   :54.80   Max.   :42.20     Max.   :80.10     
                                                    NA's   :152       
 PctUnemployed16_Over PctPrivateCoverage PctEmpPrivCoverage
 Min.   : 0.400       Min.   :22.30      Min.   :13.5      
 1st Qu.: 5.500       1st Qu.:57.20      1st Qu.:34.5      
 Median : 7.600       Median :65.10      Median :41.1      
 Mean   : 7.852       Mean   :64.35      Mean   :41.2      
 3rd Qu.: 9.700       3rd Qu.:72.10      3rd Qu.:47.7      
 Max.   :29.400       Max.   :92.30      Max.   :70.7      
                                                           
 PctPublicCoverage    PctWhite         PctBlack          PctAsian     
 Min.   :11.20     Min.   : 10.20   Min.   : 0.0000   Min.   : 0.0000  
 1st Qu.:30.90     1st Qu.: 77.30   1st Qu.: 0.6207   1st Qu.: 0.2542  
 Median :36.30     Median : 90.06   Median : 2.2476   Median : 0.5498  
 Mean   :36.25     Mean   : 83.65   Mean   : 9.1080   Mean   : 1.2540  
 3rd Qu.:41.55     3rd Qu.: 95.45   3rd Qu.:10.5097   3rd Qu.: 1.2210  
 Max.   :65.10     Max.   :100.00   Max.   :85.9478   Max.   :42.6194  
                                                                      
  PctOtherRace      PctMarriedHouseholds   BirthRate        deathRate    
 Min.   : 0.0000   Min.   :22.99        Min.   : 0.000   Min.   : 59.7  
 1st Qu.: 0.2952   1st Qu.:47.76        1st Qu.: 4.521   1st Qu.:161.2  
 Median : 0.8262   Median :51.67        Median : 5.381   Median :178.1  
 Mean   : 1.9835   Mean   :51.24        Mean   : 5.640   Mean   :178.7  
 3rd Qu.: 2.1780   3rd Qu.:55.40        3rd Qu.: 6.494   3rd Qu.:195.2  
 Max.   :41.9303   Max.   :78.08        Max.   :21.326   Max.   :362.8  
```

```r
str(Cancer, strict.width = "wrap")
```

```
'data.frame':   3047 obs. of  29 variables:
$ 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 1589 1618 1766 2051 2112 2143 2185 ...
$ 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 ...
```

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

```
print(paste0('Number of rows: ', nrow(Cancer)))
[1] "Number of rows: 3047"
print(paste0('Number of columns: ', ncol(Cancer)))
[1] "Number of columns: 29"
```

The cancer.csv file contains 29 variables (30 columns, including the first one that only contains the row numbers) and 3,047 observations. Each observation (i.e. row) includes data for a county across the US. The variables are mostly numbers and integers, except for 2 that are factors (`binnedInc` and `Geography`). Below, we explain the variables in detail and provide our assessment of the quality of the data.

Variables

- Cancer data:
  - `avgAnnCount`: The average number of new cancer cases per year per county for years 2009-2013

  - `popEst2015`: Estimated population by county 2015
- Economic status:
  - `medIncome`: Median income per county
  - `povertyPercent`: Percent of population below poverty line
  - `binnedInc`: ???
- Population age and gender:
  - `MedianAge`: Median age per county
  - `MedianAgeMale`: Median age among males per county
  - `MedianAgeFemale`: Median age among females per county
- Location:
  - `Geography`: County, State names
- Marital status:
  - `PercentMarried`: Percentage of married population
  - `PctMarriedHouseholds`: Percentage of married households per county
- Education:
  - `PctNoHS18_24`: Percentage of 18-24 year old population with no high school education
  - `PctHS18_24`: Percentage of 18-24 year old population with high school education
  - `PctSomeCol18_24`: Percentage of 18-24 year old population with some college education
  - `PctBachDeg18_24`: Percentage of 18-24 year old population with bachelor's degree
  - `PctHS25_Over`: Percentage of population above 25 years old with high school education
  - `PctBachDeg25_Over`: Percentage of population above 25 years old with bachelor's degree
- Household size:
  - `AvgHouseholdSize`: Average household size per county
- Employment status:
  - `PctEmployed16_Over`: Percentage of population above 16 years old who have jobs
  - `PctUnemployed16_Over`: Percentage of population above 16 years old with no jobs
- Health insurance coverage:
  - `PctPrivateCoverage`: Percentage of the population with private insurance coverage
  - `PctEmpPrivCoverage`: percentage of the population with employer-sponsored insurance coverage
  - `PctPublicCoverage`: Percentage of the population with public insurance coverage
- Race:
  - `PctWhite`: Percentage of white population by county
  - `PctBlack`: Percentage of African-American population by county
  - `PctAsian`: Percentage of Asian population by county
  - `PctOtherRace`: Percentage of other races by county
- Birth and death rates:
  - `BirthRate`: Birth rate per county
  - `deathRate`: Death rate per county

Evaluation of Dataset and Variables

Based on the outputs from diagnostic and summary statistics functions, as well as further univariate analysis, using relevant charts, below we describe our evaluation of the dataset and its variables. Since definition of most variables was not provided to us, our first step was to ensure understanding of what exactly such variables represent. We also evaluated the data to identify potentially erroneous values, extreme outliers and variables that might require transformation.

- **Data time frame:**: While `avgAnnCount` represents statistics for 2009-2013, the population by county is for 2015 and other variables do not have date stamps. Ideally all variables should have been from the same time period.

- **`avgAnnCount` definition:**: There is no clear definition for incidence rate per county for the `avgAnnCount` variable. Since the sum of all values is 1,847,514 and that based on Cancer.gov) data the average number of cases for 2009-2013 is 1,617,144, we assume this variable represents the actual count of new cases. Therefore, in our analysis we created a new variable called `incidenceRate` to represent the incidence rate of cancer per 100,000 people per county to be able to compare the spread of new cancer cases in different geographical regions regardless of the actual population of such regions.

```
#caclulating the total for avgAnnCount to compare with offical reports by Cancer.gov
sum(Cancer$avgAnnCount)
[1] 1847514
```

Official Cancer Statistics, 2009-2013

Year	New Cases	Deaths
2009	1,660,290	562,340
2010	1,529,560	569,490
2011	1,596,670	571,950
2012	1,638,910	577,190
2013	1,660,290	580,350

Source: Cancer.gov

```
#calculating the mean of the number of new cancer cases for years 2009-2013
#based on Cancer.gov data, in order to cofirm our assumption regarding avgAnnCount
incidence_cancer_gv <- c(1660290, 1529560, 1596670, 1638910, 1660290)
mean(incidence_cancer_gv)
[1] 1617144
```

- **Anomaly in `avgAnnCount`:** Through our assessment, we noticed that the number of new cancer cases (`avgAnnCount`) for 6 counties were greater than those counties' populations (`popEst2015`). Looking at the 6 observations, we realized that the value assigned to `popEst2015` for all these 6 counties is exactly the same number (1962.667684). In fact there are a total of 206 counties that have exactly the same average number of new cases, which is probably an error in the dataset. We decided to replace all of them with NA in our analysis.

- **`Geography`:** We checked this variable to identify potential duplicates. Since the number of unique values in this column (3,047) is equal to the total number of observations, there can not be any duplicates in this column.

```
#checking for potential dubplicates in this variable
length(unique(Cancer[["Geography"]]))
[1] 3047
```

- **`binnedInc`:** This variable has 10 levels that seem arbitrary and have different bin sizes. It is not clear why the income bins have been defined this way. As a result, we decided to ignore it in our analysis.

- **Anomaly in `MedianAge`:** The maximum `MedianAge` shows a value of 624, which is clearly a wrong number. We actually identified a total of 30 values in this column that are above 100; therefore, we will replace such values with NA in our analysis.

```
#checking the number of erroneous values
age_error = subset(Cancer, MedianAge > 100)
nrow(age_error)
[1] 30
```

- **Anomaly in `AvgHouseholdSize`:** The minimum for `AvgHouseholdSize` is 0.0221, which does not make sense, since we do not expect a household size below 1. There are 61 values in this column that are below 1, which we will replace with NA in our analysis.

```
#checking the number of erroneous values
household_error = subset(Cancer, AvgHouseholdSize < 1)
```

```
nrow(household_error)
[1] 61
```

- **`PctSomeCol18_24`:** 75% of values within this variable are NAs (2285 out 3047). Therefore, we decided to ignore this variable in our analysis.
- **`BirthRate`:** It is not clear what exactly this represents. Often, the birth rate is defined as childbirths per 1,000 people per year, but applying that to this variable would not give us the right number. For example in Los Angeles County with the population of 10,170,292, there were 124,641 live births in 2015 based on official reports, which translates into a birth rate of 12.25 (BR = (b / p) X 1,000). However, the birth rate in our data shows a value of 4.7 for this county, which is probably the ratio of women aged 15-50 years old who gave birth in 2015 as reported by TownCharts). As a result, we decided to ignore this variable in our analysis.

```
#checking the BirthRate value for Los Angeles County
Cancer[1000,'BirthRate']
[1] 4.705281
```

```
#Calculating LA County birth rate based on official figures. Formula: BR = (b ÷ p) X 1,000
124641/10170292*1000
[1] 12.2554
```

- **`deathRate`:** Based on our assessment, we believe this variable represents the number of deaths due to cancer per 100,000 population per county. For instance, we looked at the figure for Kings County, NY (173.6) and the number in our data is closer to the officially reported cancer death rate (140.3), as opposed to overall death rate (603.1). We also calculated the actual number of deaths per county (`deathRate` * `popEst2015` / 100000) and the total for these values, which is eaual to 525,347. This number is pretty close to the figure reported by Cancer.gov (589,430), further confirming our assumption regarding `deathRate`.

```
#checking the deathRate for Kings County, NY
Cancer[388, 'deathRate']
[1] 173.6
```

```
Kings County, NY statistics:
    2015 population: 2,673,000
    2015 death rate (per 100,000 population): 603.1
    2015 Cancer death rate (per 100,000 population): 140.3
```

Sources: DATA USA, NY State Dpt of Health

```
#comparing total death count in our dataset with Cancer.gov stats
Cancer$death_count <- Cancer$deathRate * Cancer$popEst2015/100000
sum(Cancer$death_count)
[1] 525347.7
```

- **`PctEmpPrivCoverage`:** We assume that the values in this variable represent a subset of values in `PctPrivateCoverage`, since the sum of these two variables in some rows is above 100.
- **Overlap between `PctPrivateCoverage` and `PctPublicCoverage`:** We assume that there is an overlap between people that have public health insurance and those with private health insurance, since the sum of `PctPrivateCoverage` and `PctPublicCoverage` in some rows is above 100. In fact, this is not uncommon among some senior citizenz that have both Medicare and a supplementary private health plan (aka. Medigap).

```
#adding up health insurance coverage variables to check for overlaps
Cancer$Pct_insured <- Cancer$PctPrivateCoverage + Cancer$PctPublicCoverage
Cancer$Pct_PersonalIsure <- Cancer$PctPrivateCoverage + Cancer$PctEmpPrivCoverage
print('Cancer$Pct_insured')
```

```
[1] "Cancer$Pct_insured"
summary(Cancer$Pct_insured)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65.40   96.25  101.30  100.61  105.80  131.70
print('Cancer$Pct_PersonalIsure')
[1] "Cancer$Pct_PersonalIsure"
summary(Cancer$Pct_PersonalIsure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.8    92.2   106.3   105.6   118.9   163.0
```

Data transformations

Based on the data evaluation mentiond before and additional analysis, we are transforming some of the valuables that have issues, as explained below.

```
#creating separarte County and State columns to enable state-wide analysis
Cancer <- Cancer %>% separate(Geography, c("County", "State"), sep = ", ", remove = FALSE)
#replacing erroneous values with NA
Cancer$MedianAge[Cancer$MedianAge > 100] <- NA
Cancer$AvgHouseholdSize[Cancer$AvgHouseholdSize < 1] <- NA
#Ramiro to add comment
bins <- seq(20000, 130000, by = 10000)
Cancer$binnedInc2 <- cut(Cancer$medIncome, breaks = bins)
# Cancer$avgAnnCount[Cancer$avgAnnCount == 1962.667684] <- NA
# Cancer$incidenceRate <- Cancer$avgAnnCount / Cancer$popEst2015 * 100000
#creating a new variable to represent actual number of deaths due to cancer per county
Cancer$death_count <- Cancer$deathRate * Cancer$popEst2015/100000
#creating a new variable to represent the percentage of population with health insurance
Cancer$Pct_insured <- Cancer$PctPrivateCoverage + Cancer$PctPublicCoverage
#Ramiro to add note
Cancer$Pct_PersonalIsure <- Cancer$PctPrivateCoverage + Cancer$PctEmpPrivCoverage
```

Univariate Analysis of Key Variables

Even though the presentation of this section takes a linear form, the actual analysis of key variables was an iterative process. The key variables were chosen based on: * Our initial hypotheses regarding variables were potentially related to `deathRate`. * The possibility that a variable/factor could be changed through interventions implemented by government health agencies to improve cancer prevention and survival. * Additional analysis that we performed to identify variables that actually had a correlation with the dependent variable.

After selecting the key variables, our approach was to focus on assessing the quality of the data (as partly explained in the Introduction) and detecting features, through univariate analysis, that are important to include when modelling the relationships of interest, such as particular features in the distributions, unusual concentrations of observations around certain values, the presence of outliers and extreme outliers, among others.

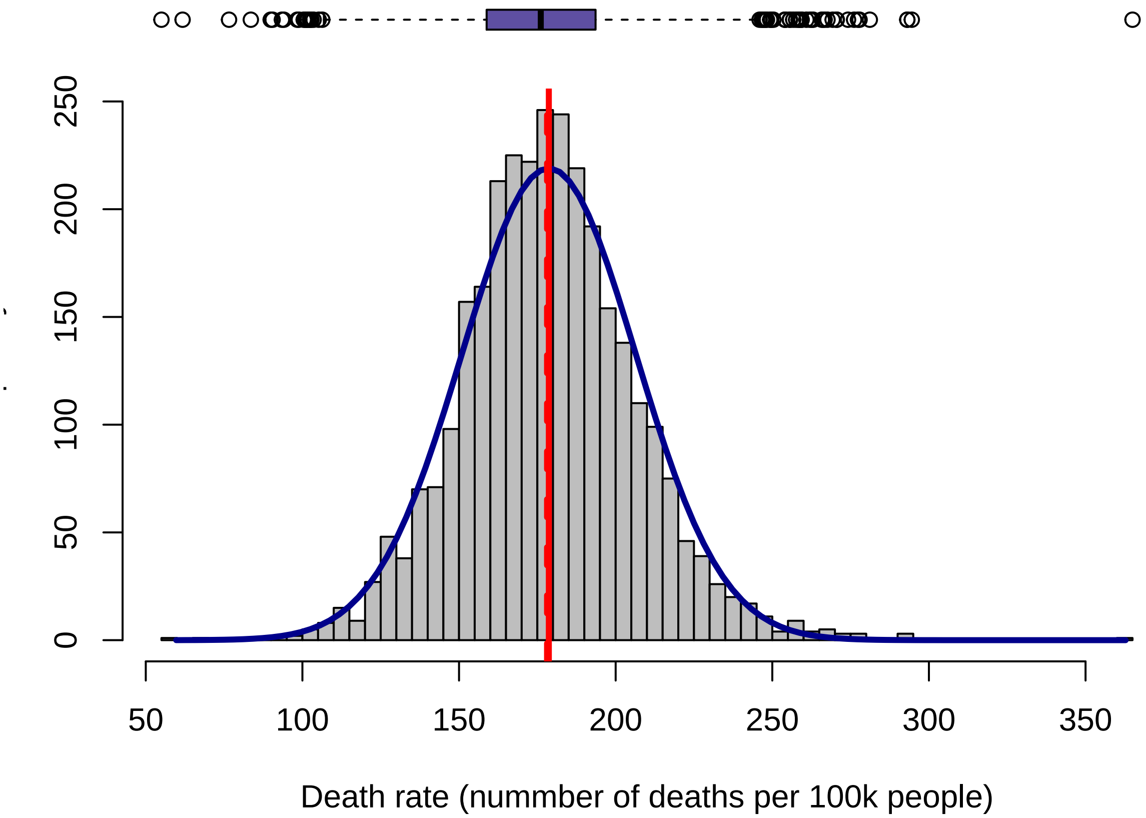
Death rate

Death rate's distribution is symmetric and bell-shaped, with a small amount of outliers at both sides of the mean (2.1% of outliers, with 0.03% of extreme outliers). However, these outliers are still within a reasonable range and do not seem to be errors in the data. Furthermore, the observation corresponding to the only extreme outlier does not look atypical based on the values of the other variables.

Finally, using both summary metrics and visualizations, we did not find any unusual concentration of observations around specific values.

```
summary(Cancer$deathRate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   59.7   161.2   178.1   178.7   195.2   362.8 
```

```
boxHist(Cancer$deathRate, "Death rate (nummber of deaths per 100k people)")
```



```
outliers.summ(Cancer, 'deathRate')
[1] "Outliers: 64 (2.1%)"
[1] "Extreme outliers: 1 (0.03%)"
```

```
Cancer[Cancer$deathRate > 300, ]
     avgAnnCount medIncome popEst2015 povertyPercent          binnedInc
1490         214     40207      15234           24.3 (37413.8, 40362.7]
     MedianAge MedianAgeMale MedianAgeFemale             Geography
1490      40.3          42.3            36.9 Union County, Florida
           County   State 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 death_count Pct_insured Pct_PersonalIsure    binnedInc2
1490     362.8    55.26895        95.4             100.6 (4e+04,5e+04]
```

Incidence (DEATILS MIGHT BE HIDDEN TO SAVE SPACE)

Looking at the frequency of unique values in 'AvgAnnCount', we found that 206 observations contain the value 1962.667684. This is very likely an error because the values in this variable should all be integers, and in some cases this value is higher than the county population.

```
incidence_freq <- data.frame(table(Cancer$avgAnnCount))
incidence_freq[incidence_freq$Freq > 20, ]
           Var1 Freq
781 1962.667684  206
```

```
table(Cancer$avgAnnCount > Cancer$popEst2015)

FALSE  TRUE
 3041     6
```

Furthermore, these values are causing the incidence rate (that we will build to be able to compare death with incidence) to have extremely large values.

Incidence rate contains 188 extremely large values (higher than 1500 cases per 100,000 people). As can be seen below, all of these values are caused by the error in AvgAnnCount.

```
Cancer$incidenceRate <- Cancer$avgAnnCount / Cancer$popEst2015 * 100000
table(Cancer$incidenceRate > 1500)

FALSE  TRUE
 2857   190
table(Cancer$incidenceRate[Cancer$avgAnnCount != 1962.667684] > 1500)

FALSE
 2841
```

Therefore, we decided to remove these "1962.667684" values and replace them with NA.

```
Cancer$avgAnnCount[Cancer$avgAnnCount == 1962.667684] <- NA
#creating new variable to represent cancer incidence per 100K people per county
Cancer$incidenceRate <- Cancer$avgAnnCount / Cancer$popEst2015 * 100000
```

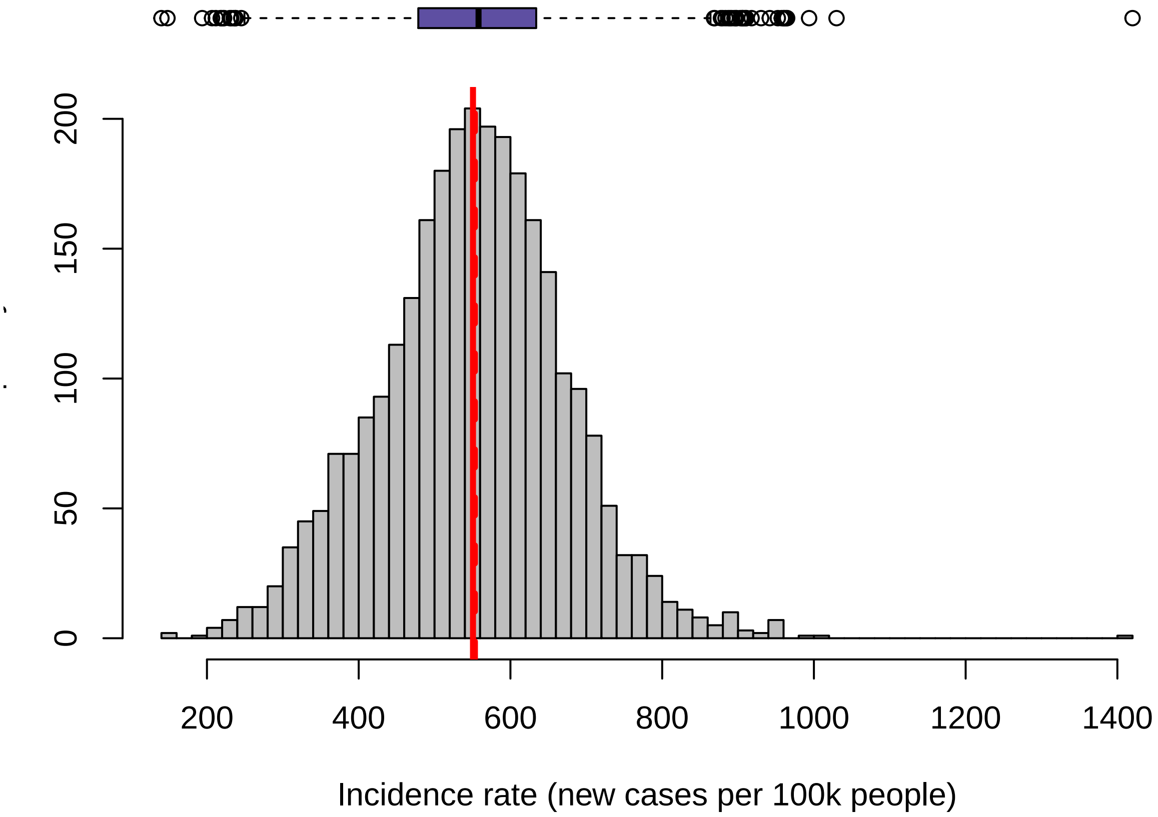
```
outliers.summ(Cancer, 'avgAnnCount')
[1] "Outliers: 334 (10.96%)"
[1] "Extreme outliers: 220 (7.22%)"
```

Incidence rate

The distribution of the incidence rate is unimodal and positively skewed, with 46 outliers and 1 extreme outlier. Since these values represent only 1.5% of observations and there is no furhter evidence that they are errors, we will keep them, but this should be taken into account when modelling the relationship between incidence and death rates.

```
summary(Cancer$incidenceRate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  140.3   474.7   553.2   550.7   628.3  1404.8     206
```

```
boxHist(Cancer$incidenceRate, "Incidence rate (new cases per 100k people)")
```



```
outliers.summ(Cancer, 'incidenceRate')
[1] "Outliers: 46 (1.51%)"
[1] "Extreme outliers: 1 (0.03%)"
```

Median income

There are two income variables available: binned income and median income. From these two, we chose median income as our key variable because it is more granular than binned income and, second, because the width of the bins in binned income seems to have been defined such that there are the same number of observations in each bin, which is not useful to observe its distribution, and the cutoffs chosen make the charts hard to read.

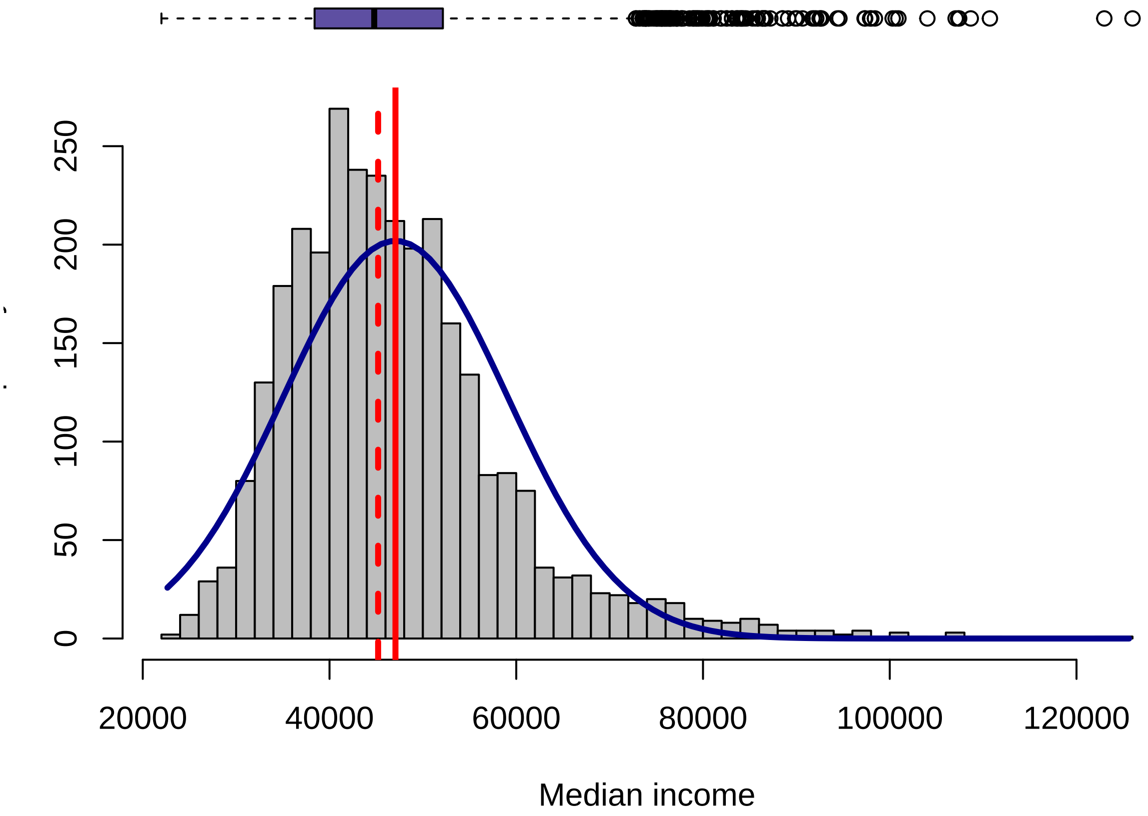
```
summary(Cancer$binnedInc)
(34218.1, 37413.8] (37413.8, 40362.7] (40362.7, 42724.4] 
               304                304                304 
  (42724.4, 45201]   (45201, 48021.6] (48021.6, 51046.4] 
               305                306                305 
(51046.4, 54545.6] (54545.6, 61494.5]  (61494.5, 125635] 
               305                306                302 
  [22640, 34218.1] 
               306 
```

Below, we can see that the median income is inded a good candidate, since it doesn't vary as much as income typically does (in this case, the difference between the minimum and maximum values is less than one order of magnitude), representing better the "average" member of each county. However, it's distribution is positively sekewed, with 64 counties that have median income greater than 80,000 USD.

```
summary(Cancer$medIncome)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22640   38882   45207   47063   52492  125635
```

```
sum(Cancer$medIncome > 80000)
[1] 64
```

```
boxHist(Cancer$medIncome, "Median income")
```



Including the 64 observations above that contribute to the positive skewness of this variable, there are still 122 outliers (around 4% of the total observations) that need to be taken into account when building the statistical model that captures the relationship between this variable and the death rate.

```
outliers.summ(Cancer, "medIncome")
[1] "Outliers: 122 (4%)"
[1] "Extreme outliers: 18 (0.59%)"
```

Given the rather large number of outliers in this variable, once might consider log transformation. However, we decided to follow the rule provided by Fox (2011), where log transformation is only likely to make a difference if its values “cover two or more orders of magnitude” (Fox, p. 128).

Education

To measure education, we have six possible candidates: ‘PctNoHS18_24’, ‘PctHS18_24’, ‘PctSomeCol18_24’, ‘PctBachDeg18_24’, ‘PctHS25_Over’ and ‘PctBachDeg25_Over’, that can be divided into two groups: 18-24 and ‘25 and above’ years old. Our initial hypothesis was that the second group should have a stronger correlation with death rate. We validated this hypothesis with the correlations table shown below, that shows the only variable that has correlation with deathRate is `PctBachDeg18_24` (although this correlation is not strong, -0.31). Instead, as expected, the two '25 and above' education variables have a much greater correlation with `deathRate`.

Therefore, we will focus on these two variables for further analyses on education.

```
cor(Cancer[, names(Cancer) %in%
           c('PctNoHS18_24', 'PctHS18_24', 'PctSomeCol18_24', 'PctBachDeg18_24',
             'PctHS25_Over', 'PctBachDeg25_Over', 'deathRate')], use = 'complete.obs')[7, ]
     PctNoHS18_24        PctHS18_24   PctSomeCol18_24   PctBachDeg18_24
        0.1219703         0.2665730        -0.1886877        -0.3140130
     PctHS25_Over PctBachDeg25_Over         deathRate
        0.4182411        -0.4717962         1.0000000
```

We also validated that education variables within each group are mutually exclusive, by making sure that they add up to 100% for all observations that have complete data. The total values range between 99.9 and 100.1, where the small variations around 100 are most probably due to rounding.

We can only test this with the 18-24 group since the 25_over group is missing two variables that capture 'no high school' and 'some college'. However, it is reasonable to assume that the same definition is applied to our group of interest (25_over).

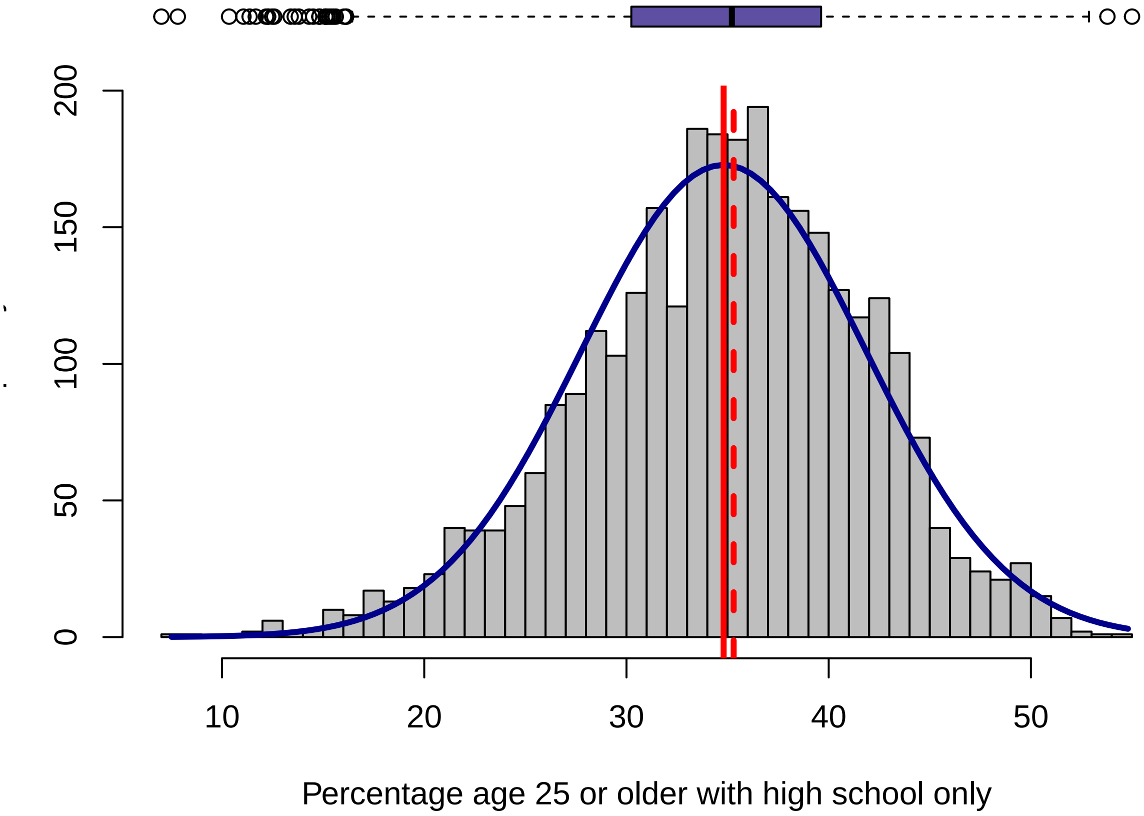
```
educ.18.24 <- c('PctNoHS18_24', 'PctHS18_24', 'PctSomeCol18_24', 'PctBachDeg18_24')
educ.df <- subset(Cancer, select = educ.18.24)
educ.complete <- complete.cases(educ.df)
sum.pct.freq <- data.frame(table(rowSums(educ.df[educ.complete, ], na.rm = TRUE)))
names(sum.pct.freq) <- c("Sum", "Frequency")
sum.pct.freq
    Sum Frequency
1  99.9       127
2   100       518
3 100.1       117
```

PctHS25_over

Values in `PctHS25` are within a reasonable range (7 to 55%) and there doesn't seem to be an unusual concentration of observations around certain values. Also, the disribution of this variable is unimodal and negatively skewed. However, it only contains 31 outliers (1% of observations) and there are no extreme outliers. Furthermore, there is no indication that these outliers are errors, so we decided to keep them.

```
summary(Cancer$PctHS25_Over)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.50   30.40   35.30   34.80   39.65   54.80
```

```
boxHist(Cancer$PctHS25_Over, "Percentage age 25 or older with high school only")
```



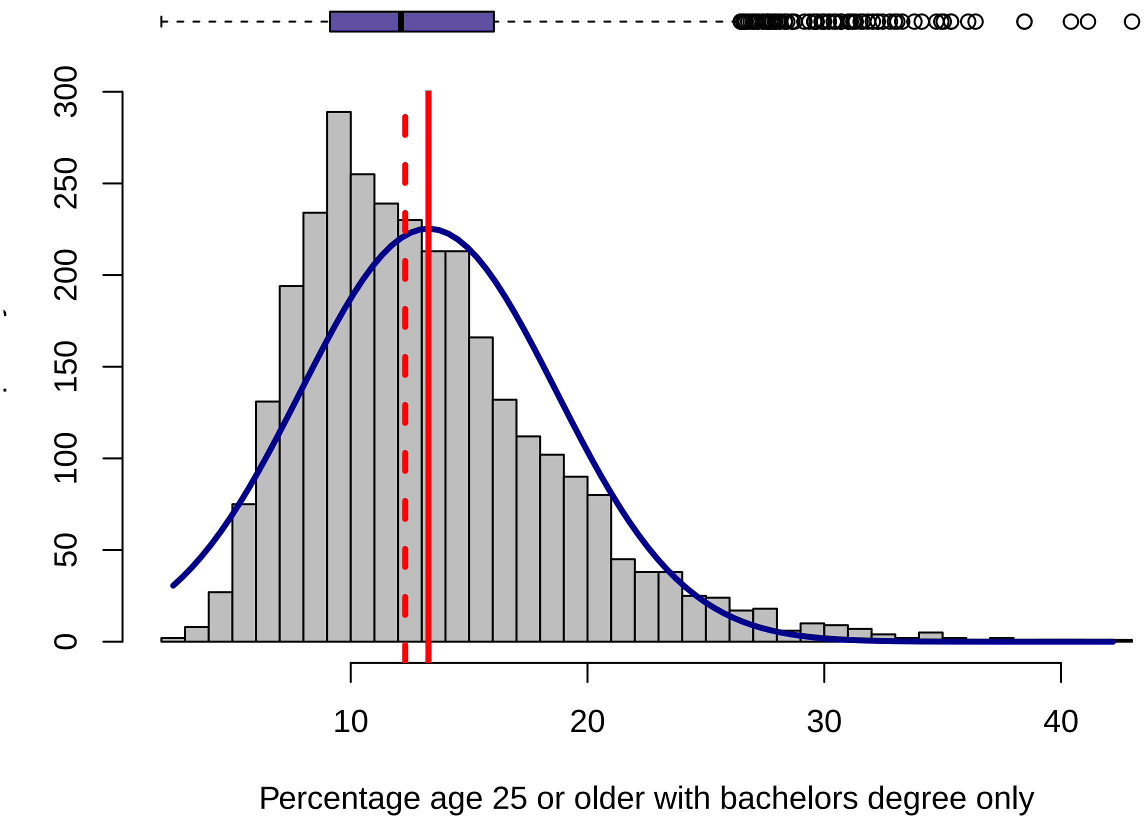
```
outliers.summ(Cancer, "PctHS25_Over")
[1] "Outliers: 31 (1.02%)"
[1] "Extreme outliers: 0 (0%)"
```

PctBachDeg25_Over

Values in `PctHS25_Over` are within a reasonable range (7% to 55%) and there doesn't seem to be an unusual concentration of observations around certain values. The disribution of this variable is unimodal and positively skewed. It contains 82 outliers (2.7% of observations) all of which at are at the right side of the mean. Of these 82 outliers, only 5 are extreme outliers, that we will keep in the dataset, since there are no indications that they are errors.

```
summary(Cancer$PctBachDeg25_Over)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.50    9.40   12.30   13.28   16.10   42.20
```

```
boxHist(Cancer$PctBachDeg25_Over, "Percentage age 25 or older with bachelors degree only")
```



Extreme outliers

```
outliers.summ(Cancer, "PctBachDeg25_Over")
[1] "Outliers: 82 (2.69%)"
[1] "Extreme outliers: 5 (0.16%)"
```

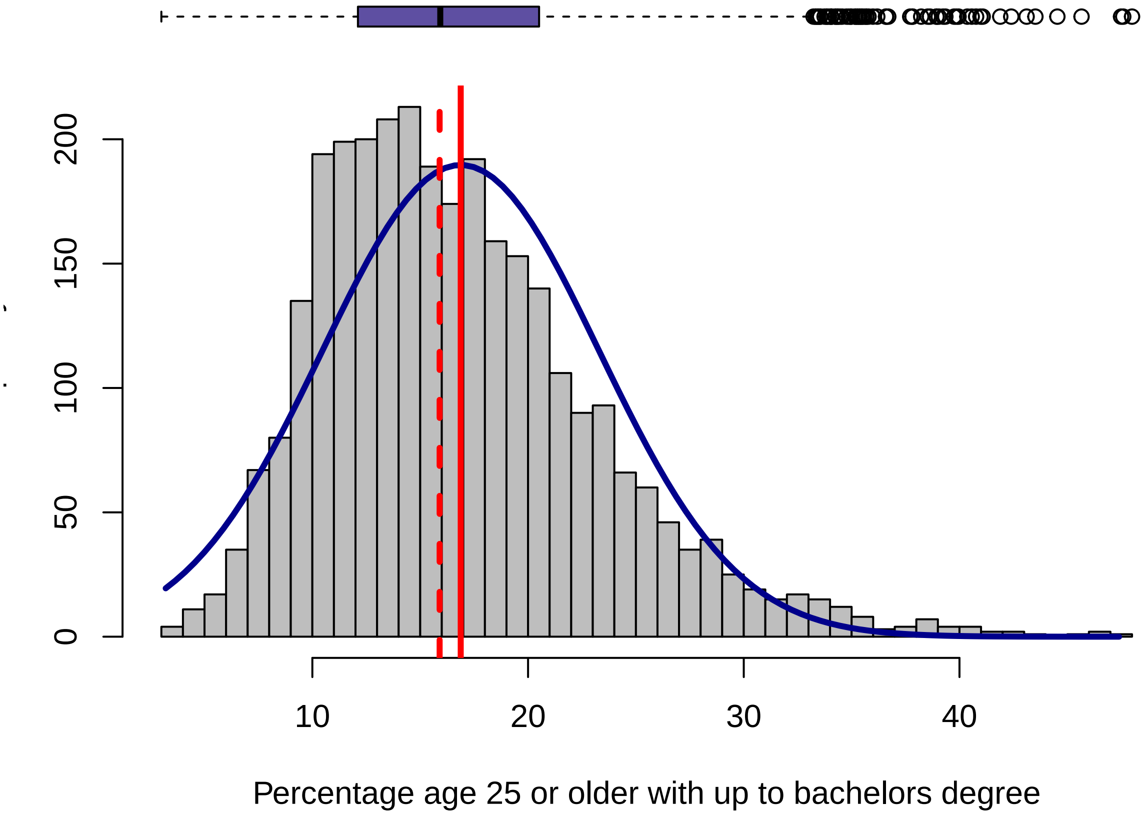
Poverty percent

The distribution of `povertyPercent` is unimodal and positively skewed. This is reflected by the fact that all outliers are at the right of the mean. Taking a deeper dive into the outliers, we found that only 3 are extreme while 66 are mild. For this reason, and because we did not find other indication that the outliers or other values were errors, we will keep all data from this variable.

However, when modeling the relationship of interest, we should take into account that the distribution of this variable is not normal and it might need transformation if the model used requires it.

```
summary(Cancer$povertyPercent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.20   12.15   15.90   16.88   20.40   47.40
```

```
boxHist(Cancer$povertyPercent, "Percentage age 25 or older with up to bachelors degree")
```



Extreme outliers

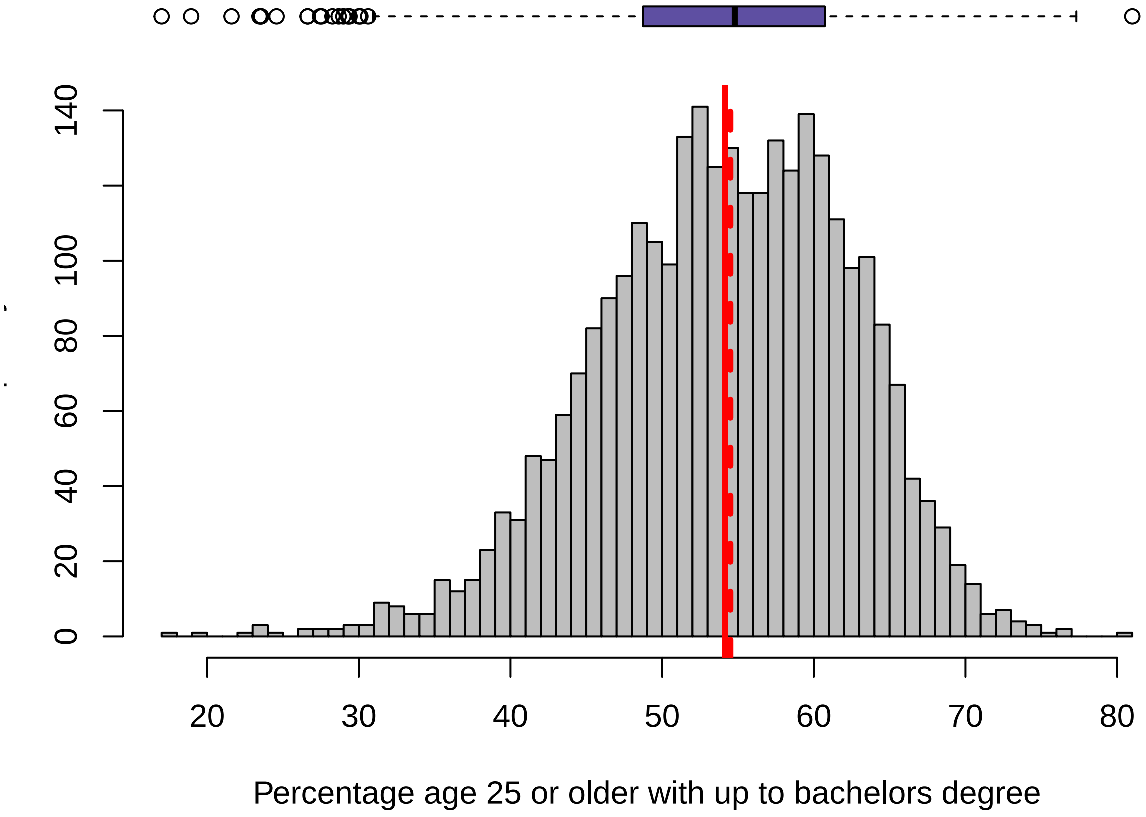
```
outliers.summ(Cancer, "povertyPercent")
[1] "Outliers: 69 (2.26%)"
[1] "Extreme outliers: 3 (0.1%)"
```

Percentage employed (16 or over)

The distribution of `PctEmployed16_Over` is unimodal and negatively skewed. There are no extreme outliers and 20 mild outliers (0.7% of observations). For this reason, and because we did not find any indication that the outliers or other values were errors, we will keep all the values within this variable.

```
summary(Cancer$PctEmployed16_Over)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  17.60   48.60   54.50   54.15   60.30   80.10     152 
```

```
boxHist(Cancer$PctEmployed16_Over, "Percentage age 25 or older with up to bachelors degree")
```



Extreme outliers

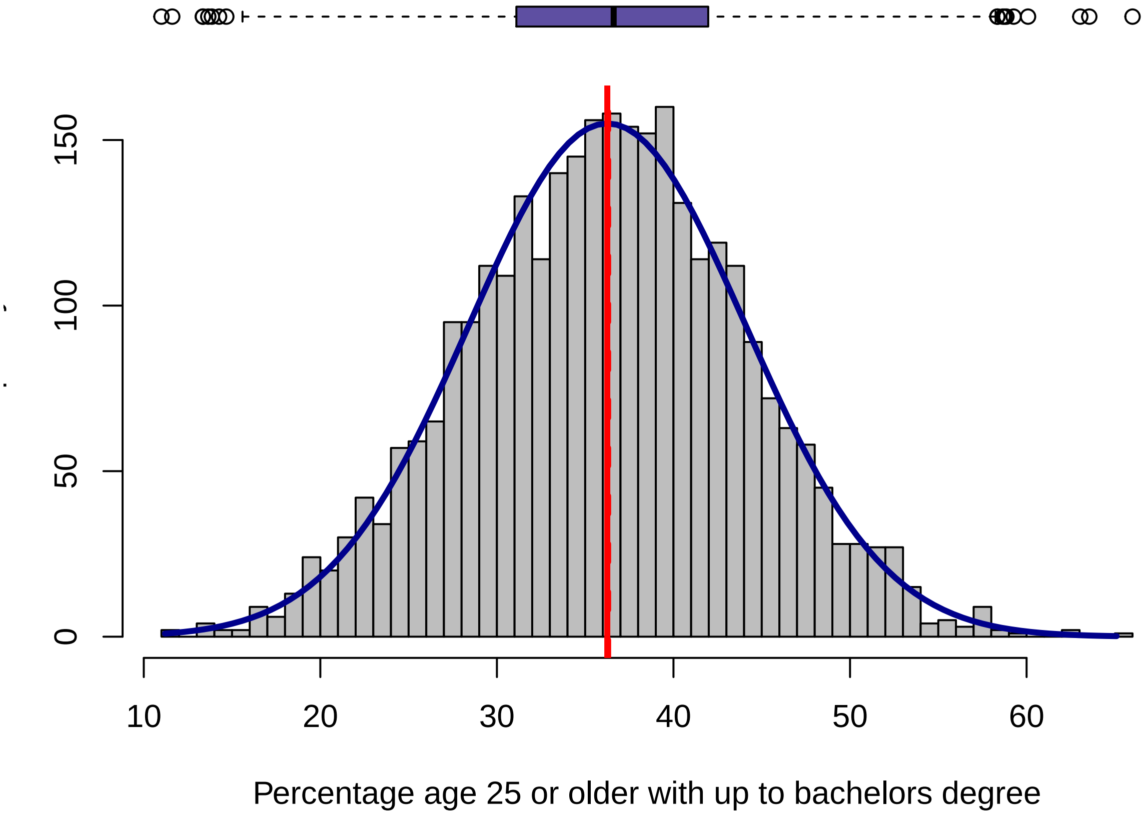
```
outliers.summ(Cancer, "PctEmployed16_Over")
[1] "Outliers: 20 (0.66%)"
[1] "Extreme outliers: 0 (0%)"
```

Percentage with public coverage

The distribution of `PctPublicCoverage` is unimodal and symmetric, with no extreme outliers and only 18 mild outliers (0.6% of observations). For this reason, and because we did not find any indication that the outliers or other values were errors, we will keep all data from this variable. There are also no other particular features from this variables that grant further warnings in modelling the relationship with `deathRate`.

```
summary(Cancer$PctPublicCoverage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  11.20   30.90   36.30   36.25   41.55   65.10 
```

```
boxHist(Cancer$PctPublicCoverage, "Percentage age 25 or older with up to bachelors degree")
```



Extreme outliers

```
outliers.summ(Cancer, "PctPublicCoverage")
[1] "Outliers: 18 (0.59%)"
[1] "Extreme outliers: 0 (0%)"
```

Analysis of Key Relationships

Education

As explained above, guided by our hypothesis that the education of the '25 and over' years old group should have a much stronger relationship with deathRate than the '18-24' years old group, which was supported by the correlations between these variables, we will focus on the former group.

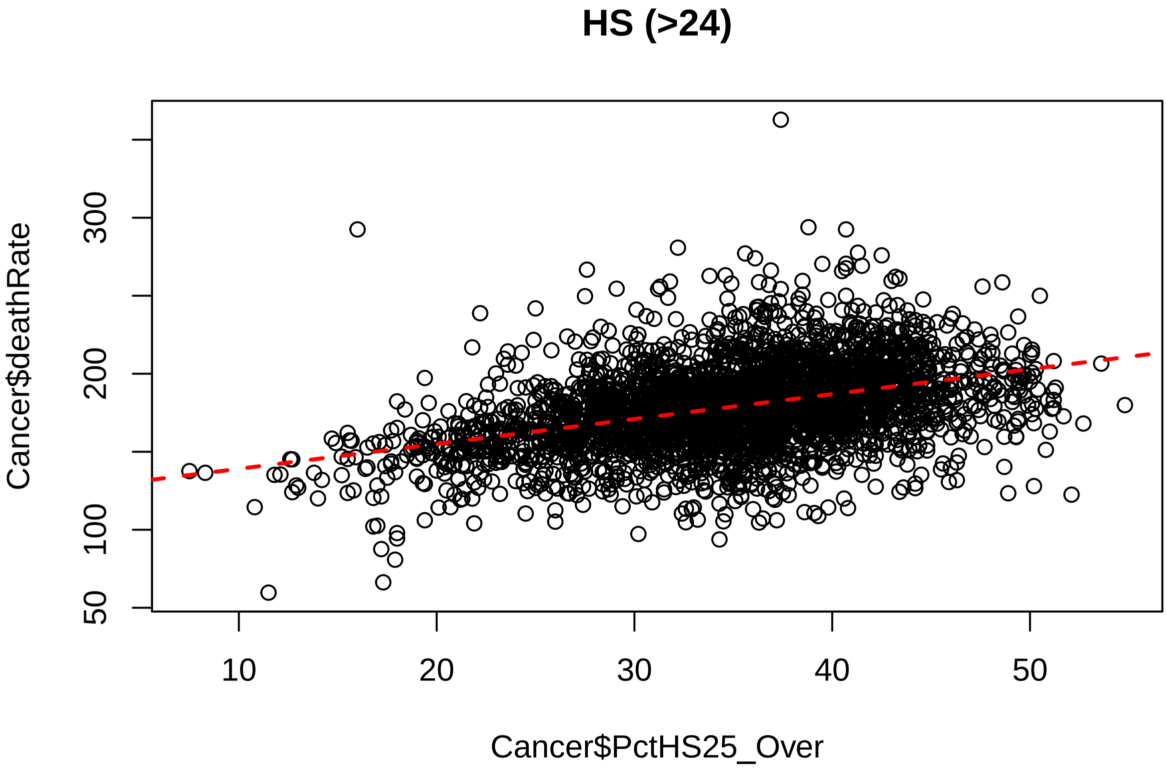
PctHS25_over

A correlation of 0.4 between `PctHS25_over` and `deathRate` indicates that there is indeed a relationship between these variables, which is further indicated by plotting them together in a scatterplot, that shows that higher values of percentage of population with only high school tend to be associated to higher death rates (this is also reflected in the regression line added to the scatterplot).

Such relationship is not unexpected, since it indicates that a higher concentration of people with low education levels may have poorer health habits and limited access to medical services. However, both of these variables could be affeccted by `MedianAge` in the same direction: older populations might have lower levels of higher education and higher rates of death.

```
cor(Cancer$deathRate, Cancer$PctHS25_Over)
[1] 0.4045891
```

```r
plot(Cancer$PctHS25_Over, Cancer$deathRate, main = "HS (>24)")
abline(lm(Cancer$deathRate ~ Cancer$PctHS25_Over), lty = 'dashed', lwd = 2, col = 'red')
```



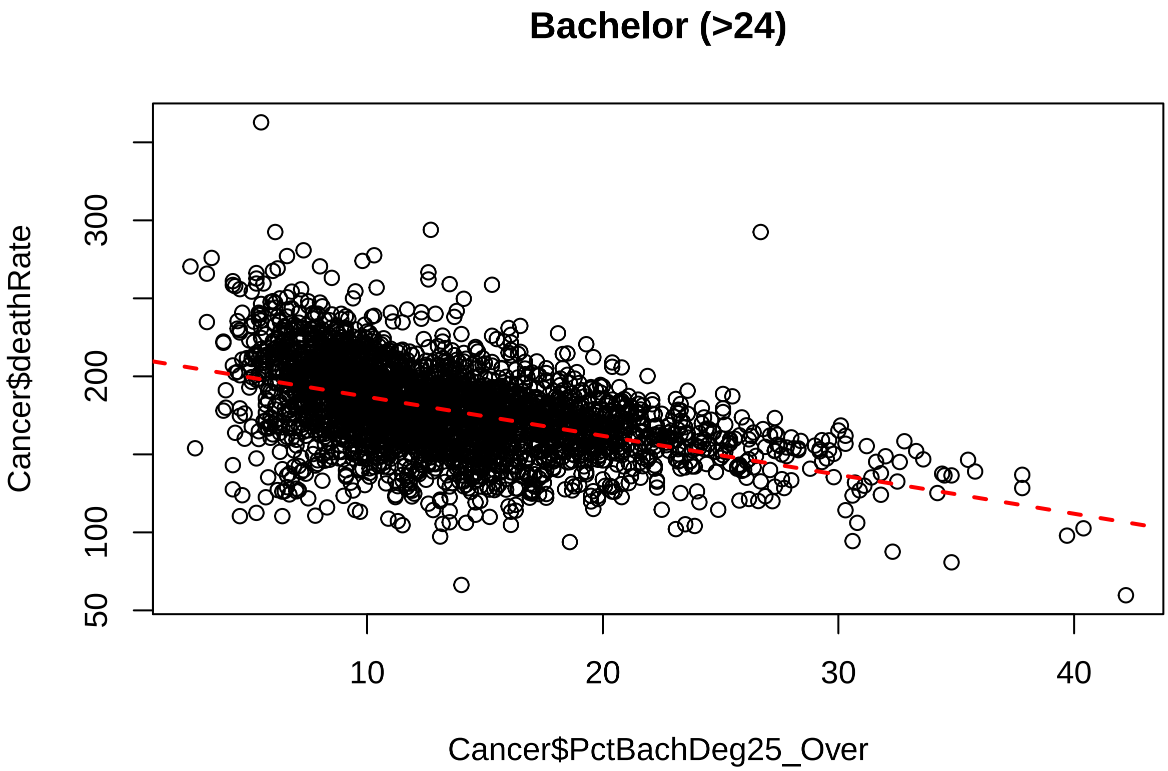
PctBachDeg25_Over

A correlation of −0.48 indicates that there is relationship between `PctBachDeg25_over` and `deathRate`, which is further supported by plotting these variables in a scatterplot, where it can be seen that a higher percentage of people with bachelor's degree is associated with lower levels of death rates. This relationship is also supported by the regression line included in the scatterplot.

This is also not an unusual relationship, since higher levels of education might be linked to better health habits and access to health services. However, and following the same reasoning than `PctHS25_over`, the relationship between these two variables may be confounded by `MedianAge`, although it is not clear in which direction this effect might go. Therefore, it will also be necessary to explore the effect of `MedianAge` in the following section.

```r
cor(Cancer$deathRate, Cancer$PctBachDeg25_Over)
[1] -0.4854773
```

```r
plot(Cancer$PctBachDeg25_Over, Cancer$deathRate, main = "Bachelor (>24)")
abline(lm(Cancer$deathRate ~ Cancer$PctBachDeg25_Over), lty = 'dashed', lwd = 2, col = 'red')
```



Analysis of Secondary Effects

Throughout the analyses above, we began to identify that some of the relationships found between `deathRate` and other variables may not only be capturing the direct relationship between these variables, but also those of additional variable(s) that may be impacting both. To further assess this systematically, the following network visualization shows the variables that have a correlation higher than 0.4, where each node represents a different variable and each vertex indicates the strength of the relationship between the variables connected.

[secondary_analysis]

(secondary_analysis.png)

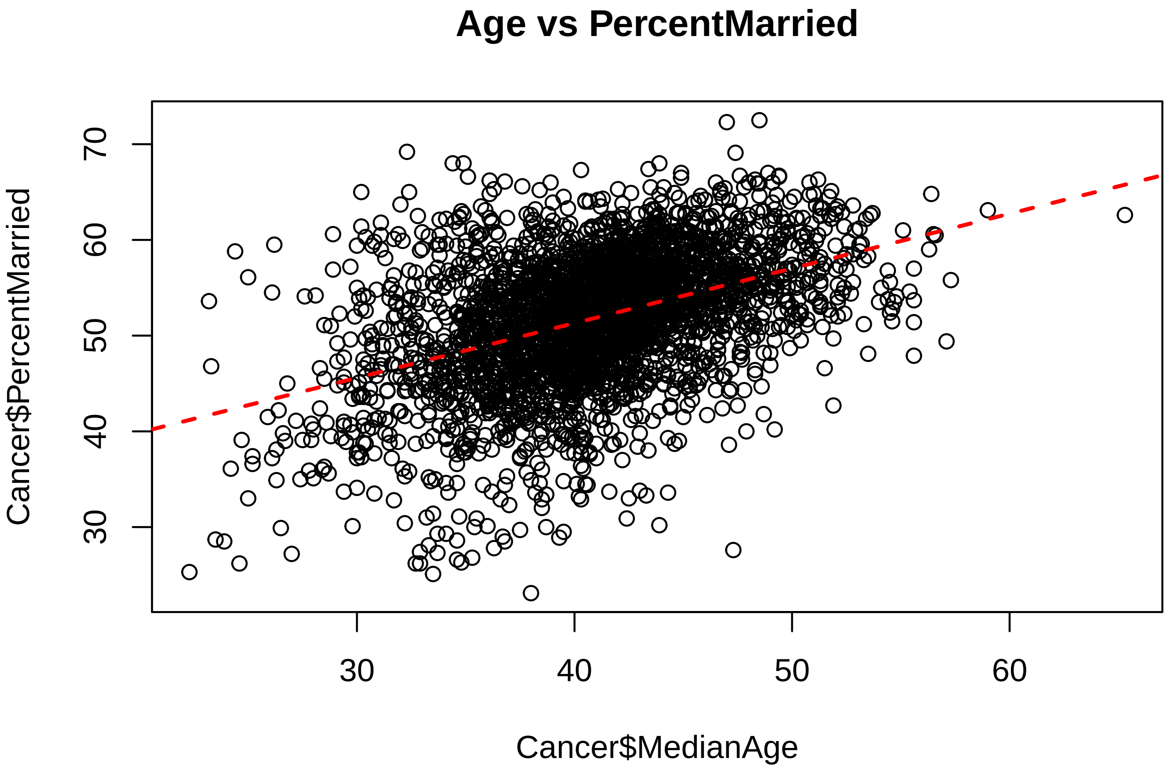
Age and family/householdd

`PercentMarried` has a (weak) relation and `AvgHouseholdSize` has a moderate relation with `MedianAge`. Based on these results, we explored this relationship further.

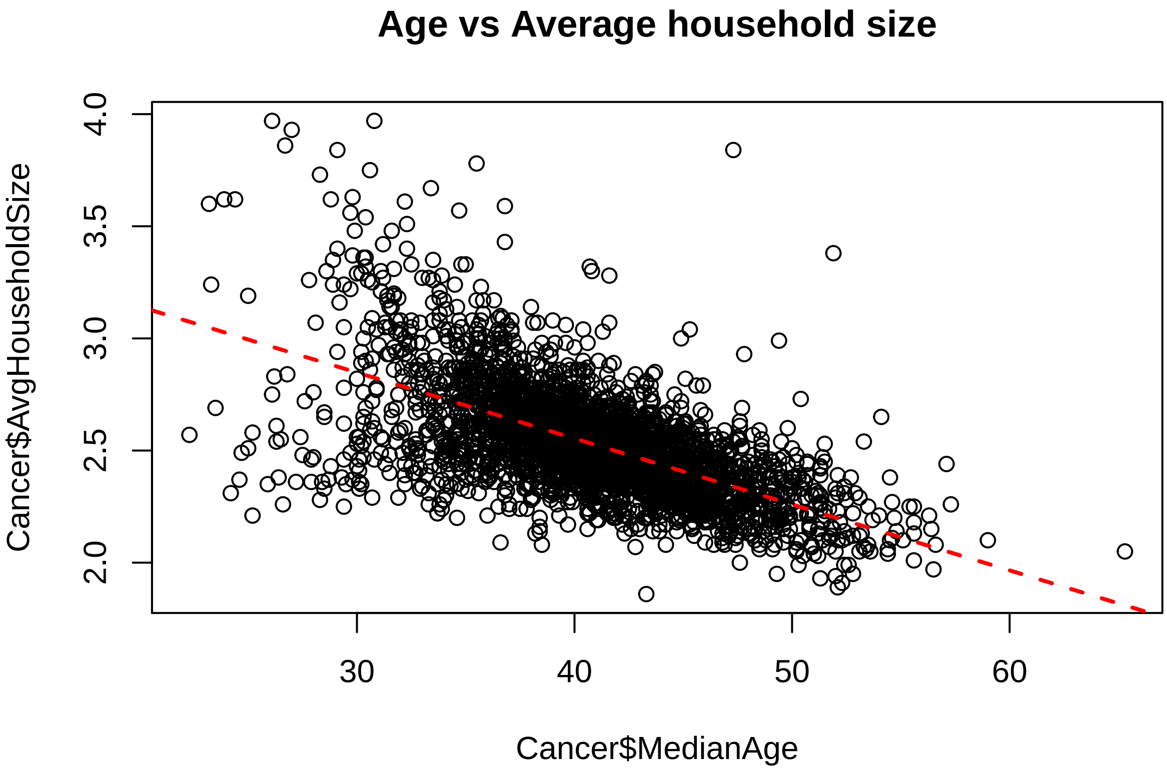
```
cor(subset(Cancer,
           select = c("MedianAge", "AvgHouseholdSize", "PercentMarried")),
    use = "pairwise.complete.obs")[1, ]
       MedianAge AvgHouseholdSize   PercentMarried
       1.0000000       -0.6145498        0.4301158
```

Both the scatterplots below and the regression lines imposed on them provide further support that there is indeed a relationship between these two variabes and `MedianAge`, indicating that `MedianAge` may confound the relationship between these two and `deathRate`. Therefore, this should be taken into account when modelling the relationships of interest, in order to isolate the effect of the family variables on the death rate.

```r
plot(Cancer$MedianAge, Cancer$PercentMarried, main = "Age vs PercentMarried")
abline(lm(PercentMarried ~ MedianAge, data = Cancer), lty = 'dashed', lwd = 2, col = 'red')
```



```r
plot(Cancer$MedianAge, Cancer$AvgHouseholdSize, main = "Age vs Average household size")
abline(lm(AvgHouseholdSize ~ MedianAge, data = Cancer), lty = 'dashed', lwd = 2, col = 'red')
```

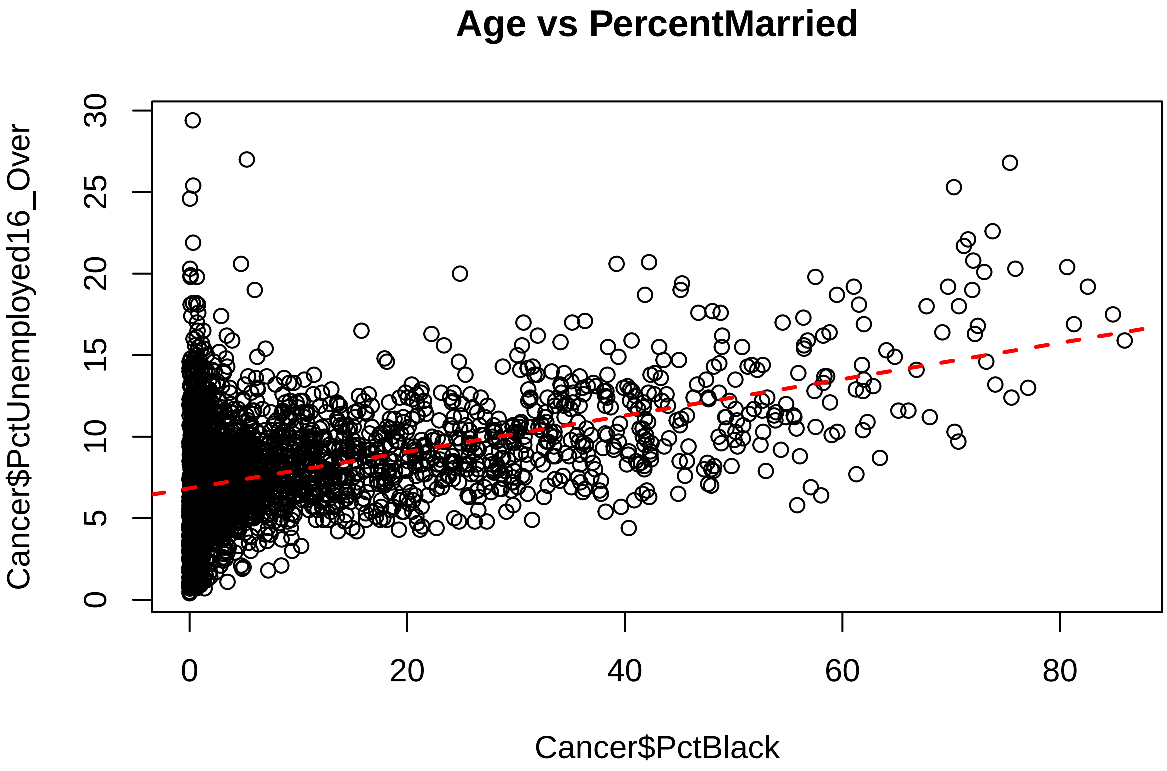


Black population and employment

Correlation analysis shows that there is a relationship between the percentage of black population and employment, which is further confirmed both by a visual inspection of the scatterplot and the linear regression line charted in this plot. Since employment es related to `deathRate`, its correlation with `PctBlack` may indicate that this variable may be confounding the relationship of interest and thus further modeling needs to take this into account, to isolate the effect of unemployment on death rate.

```
cor(subset(Cancer, select = c("PctBlack", "PctUnemployed16_Over")),
    use = "pairwise.complete.obs")[, 1]
           PctBlack PctUnemployed16_Over
          1.0000000            0.4692731
```

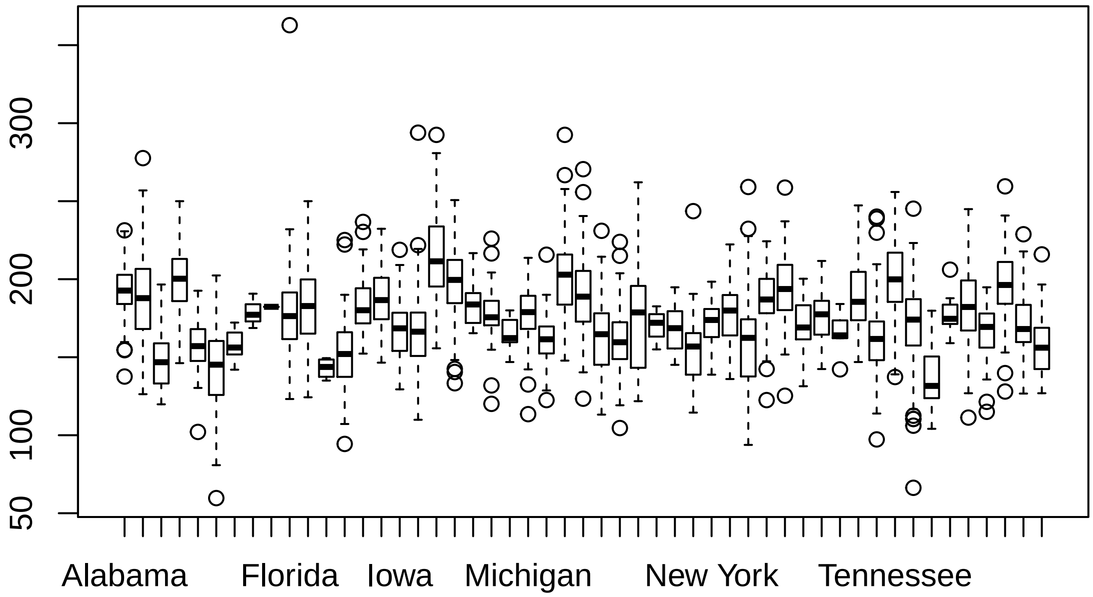
```
plot(Cancer$PctBlack, Cancer$PctUnemployed16_Over, main = "Age vs PercentMarried")
abline(lm(PctUnemployed16_Over ~ PctBlack, data = Cancer), lty = 'dashed', lwd = 2, col = 'red')
```



State

A boxplot containing different location measures of `deathRate` by `State` shows that these values vary measures vary significantly across state. Since `State` may be capturing several state-level characteristics that may in turn affect other variables that have a relation with `deathRate`, it is recommended to include state-level effects when modeling the relation of interest, to control for confounding these state-level factors.

```
boxplot(Cancer$deathRate ~ Cancer$State)
```


50
100
200
300
Alabama
Florida
Iowa
Michigan
New York
Tennessee