An Exploratory Analysis of Cancer Incidence and Mortality

Identifying High-Risk Communities to Improve Survival

Ramiro Cadavid, Pri Nonis, Payman Roghani

September 24, 2018

Contents

1. 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 Dataset

```
Cancer <- read.csv('cancer.csv', header = T, as.is = T, row.names = 1) # load the cancer dataset
```

```
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 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`).

Taking a first look at the data, we wanted to do some spasial data exploration. Cancer affects a significant population and the effect is felt mostly uniformly across the country.

Geographical Analysis of the Dataset

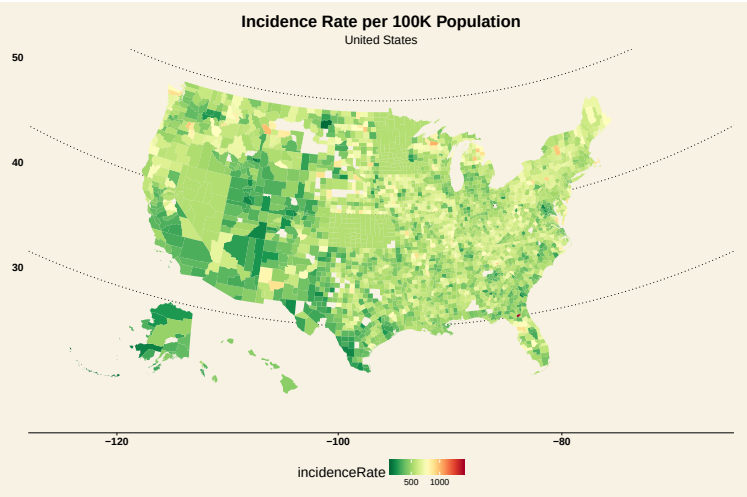
Before we could do this we needed to augment our dataset with additional geocoordinate details. We were able to obtain that data from the website of the Census Beauru at census.gov/geo/maps-data. After some minor data manipulations, we were able to join the two datasets using the **Geography** variable.

```
source("maps.R")
map_setup()
```

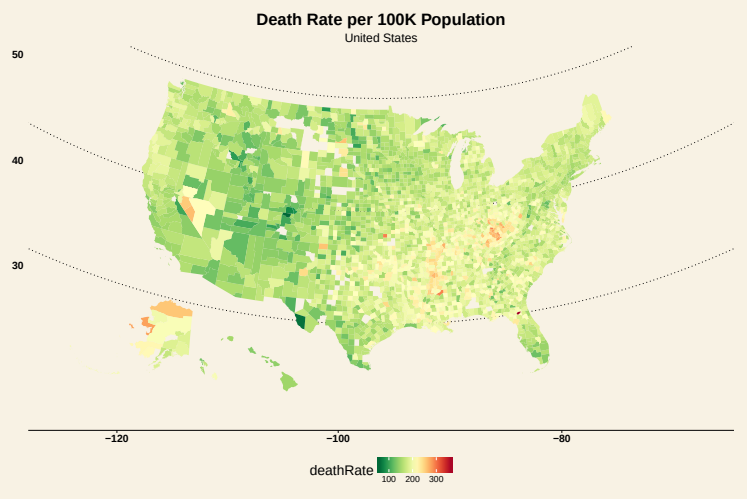
```
## [1] "Loaded Cancer Map Data"
```

This allowed us to visualize and explore the Cancer dataset:

```
map_us_incidenceRate()
```

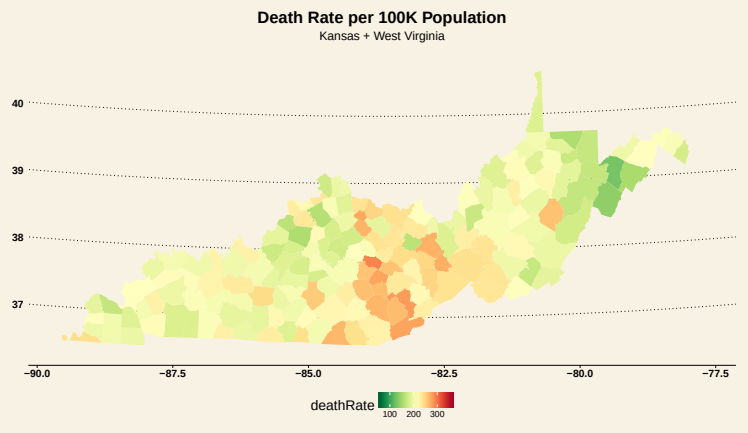


```
map_us_deathRate()
```

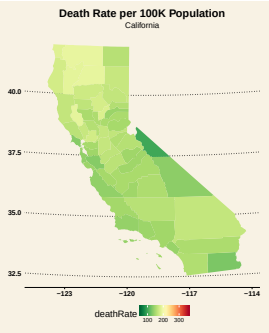


However, we did find some outliers, espcially in **Kentucky** and **West Virgiana** that is seeing a dispropotional effect, compared to other states such as California.

```
map_wv_ky_deathRate()
```




```
map_ca_deathRate()
```



We also discovered that the Cancer dataset was missing data from 93 counties belonging to the continental United States. These counties appear as blank polygons in our maps.

Below, we explain the variables in detail and provide our assessment of the quality of the data.

data on smoking and obesity and other cancer risk factors could've been very helpful

Overview of 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 regarldess 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
Source: [Cancer.gov](https://www.cancer.gov/)
```

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

```
#calculating the mean of the number of new cancer cases for years 2009-2013, based on Cancer.gov data,
incidence_cancer <- c(1660290, 1529560, 1596670, 1638910, 1660290)
mean(incidence_cancer)
[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.

```
#cheking the number of observatios, where new case count is greater than the population
sum(Cancer$avgAnnCount > Cancer$popEst2015, na.rm = TRUE)
[1] 6
```

```
Cancer$avgAnnCount[Cancer$avgAnnCount == 1962.667684] <- NA #removing the potentially erroneous number
Cancer$incidenceRate <- Cancer$avgAnnCount / Cancer$popEst2015 * 100000 #creating a new variable: new c
```

- **`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.

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

```
# we can recover these numbers #line 301
#Cancer$MedianAge[Cancer$MedianAge > 100] <- NA #replacing erroneous values with NA
```

- **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.

```
household_error = subset(Cancer, AvgHouseholdSize < 1) #checking the number of erroneous values
nrow(household_error)
[1] 61
```

```
# we can recover these numbers #line 301
#Cancer$AvgHouseholdSize[Cancer$AvgHouseholdSize < 1] = NA #replacing erroneous values with NA
```

- **`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](https://datausa.io/), [NY State Dpt of Health](https://www.health.ny.gov/)
```

```
#comparing total death count in our dataset with official stats reported by health authorities
death_count <- Cancer$deathRate * Cancer$popEst2015/100000
```

```
sum(Cancer$death_count)
[1] 0
```

- **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 makes sence of such variables and check for overlaps
Pct_insured <- Cancer$PctPrivateCoverage + Cancer$PctPublicCoverage
Pct_PersonalIsure <- Cancer$PctPrivateCoverage + Cancer$PctEmpPrivCoverage
print('Pct_insured')
[1] "Pct_insured"
summary(Pct_insured)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65.40   96.25  101.30  100.61  105.80  131.70
print('Pct_PersonalIsure')
[1] "Pct_PersonalIsure"
summary(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.

```
Cancer$color <- '#098154'
Cancer$color[Cancer$MedianAge>100] <- '#c72e29' # highlight anomalous values
Cancer$color[Cancer$AvgHouseholdSize<1] <- '#c72e29' # highlight anomalous values
Cancer$color[Cancer$avgAnnCount==1962.667684] <- '#c72e29' # highlight anomalous values
```

Breakup Geography

We will use the State variable to do more aggregate analysis.

```
Cancer <- separate(Cancer, col = Geography, into = c("County","State"), sep = ", ", remove = FALSE)
```

Median Age Corrections

The Median Age variable contained several observations that were larger than 100. As we have determined, based on the median, that this variable is given in terms of years these obeservations did not make sense. However, looking at the clustering pattern it was reasonable to assume that these specific values were given in terms of months and not years.

```
Cancer$MedianAge[Cancer$MedianAge>100] <- Cancer$MedianAge[Cancer$MedianAge>100] / 12 # correct median
```

Average Household Size Corrections

The Average Household Size for several observation were below 1, i.e. more houses in that county that people. Looking closer to that group of outliers it looked like they were off by a factor of 100 from the mean and median of the variable. Multiplying these observations by 100 brought them into a acceptable range within the first SD.

```
Cancer$AvgHouseholdSize[Cancer$AvgHouseholdSize<1] <- Cancer$AvgHouseholdSize[Cancer$AvgHouseholdSize<1
```

Average Annual Count Corrections

The Average Annual Count variable which designate the number of new cancer cases within the county contained a significant number of values that seems to be some data inputing error, they were all 1962.667684. This value is too large to be resonably considered a valid incidence rate. These datapoints were from only 3 states. So we explored several approaches to correct these values.

- **Keep the Anomalous Values** This would however change the sample size of variables we want to compare with each other.
- **Remove Entire Observation with the Anomalous Values**. This would however remove a large percentage of our dataset and significantly weaken our analysis.
- **Remove Just the Anomalous Values** This would however change the sample size of variables we want to compare with each other.
- **Replae the Anomalous Values by Imputation** Based on the statistical relationship between the county population and the incidence count, we can replace the anomalous values. First by finding the mean of the avgAnnCount without the 1962.667684 value. Then we can use this mean to multiply the popEst2015 to replace all values of 1962.667684.

Choice 1 - Not Evaluvated

```
incidenceMean <- mean(Cancer$avgAnnCount[Cancer$avgAnnCount!=1962.667684] / Cancer$popEst2015[Cancer$av
Cancer$avgAnnCount[Cancer$avgAnnCount==1962.667684] <- incidenceMean * Cancer$popEst2015[Cancer$avgAnnC
Cancer$incidenceRate <- Cancer$avgAnnCount / Cancer$popEst2015 * 100000 # new cancer cases per 100,000
```

Choice 2 - Final Choice

```
Cancer$avgAnnCount[Cancer$avgAnnCount==1962.667684] <- NA # remove just the anomalous values
Cancer$incidenceRate <- Cancer$avgAnnCount / Cancer$popEst2015 * 100000 # new cancer cases per 100,000
```

2. 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

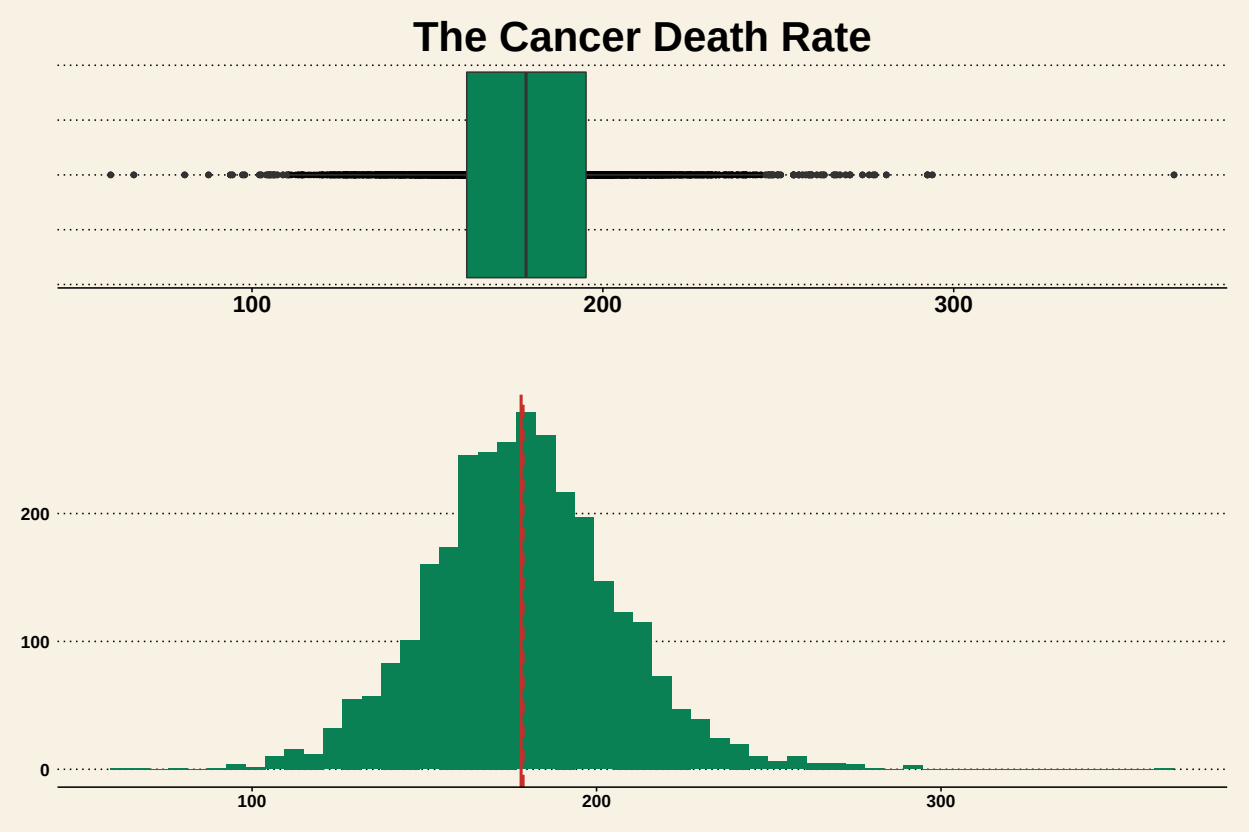
Death rate's distribution is symmetric and bell-shaped, with a small amount of outliers at both sides of the mean (2.1% of outliers, with 0.03% of extreme outliers). However, these outliers are still within a reasonable range and do not seem to be errors in the data. Furthermore, the observation corresponding 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, "The Cancer 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 incidenceRate   color
```

```
## 1490     362.8      1404.753 #098154
```

Incidence

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, ]
```

```
## [1] Var1 Freq
## <0 rows> (or 0-length row.names)
```

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

```
##
## FALSE
##  2841
```

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
##  2841
```

```
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
Cancer$incidenceRate <- Cancer$avgAnnCount / Cancer$popEst2015 * 100000
```

```
outliers.summ(Cancer, 'avgAnnCount')
```

```
## [1] "Outliers: 334 (10.96%)"
## [1] "Extreme outliers: 220 (7.22%)"
```

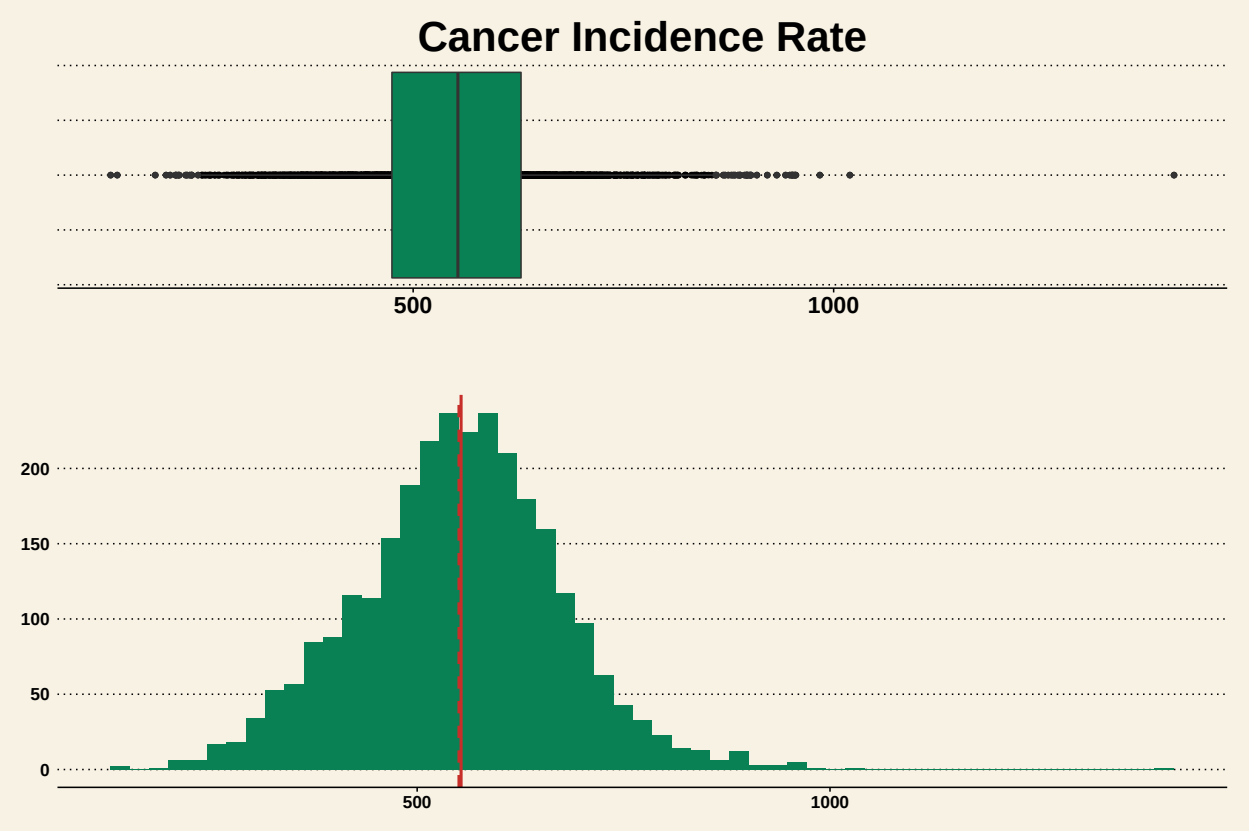
Incidence Rate

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, they will be kept, but 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, "Cancer Incidence Rate", "New Diagnosed 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 binned income seem to have been defined to have a similar 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)
```

```
##    Length     Class      Mode
##      3047 character character
```

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, having 64 counties where the median income is higher than 80,000 USD.

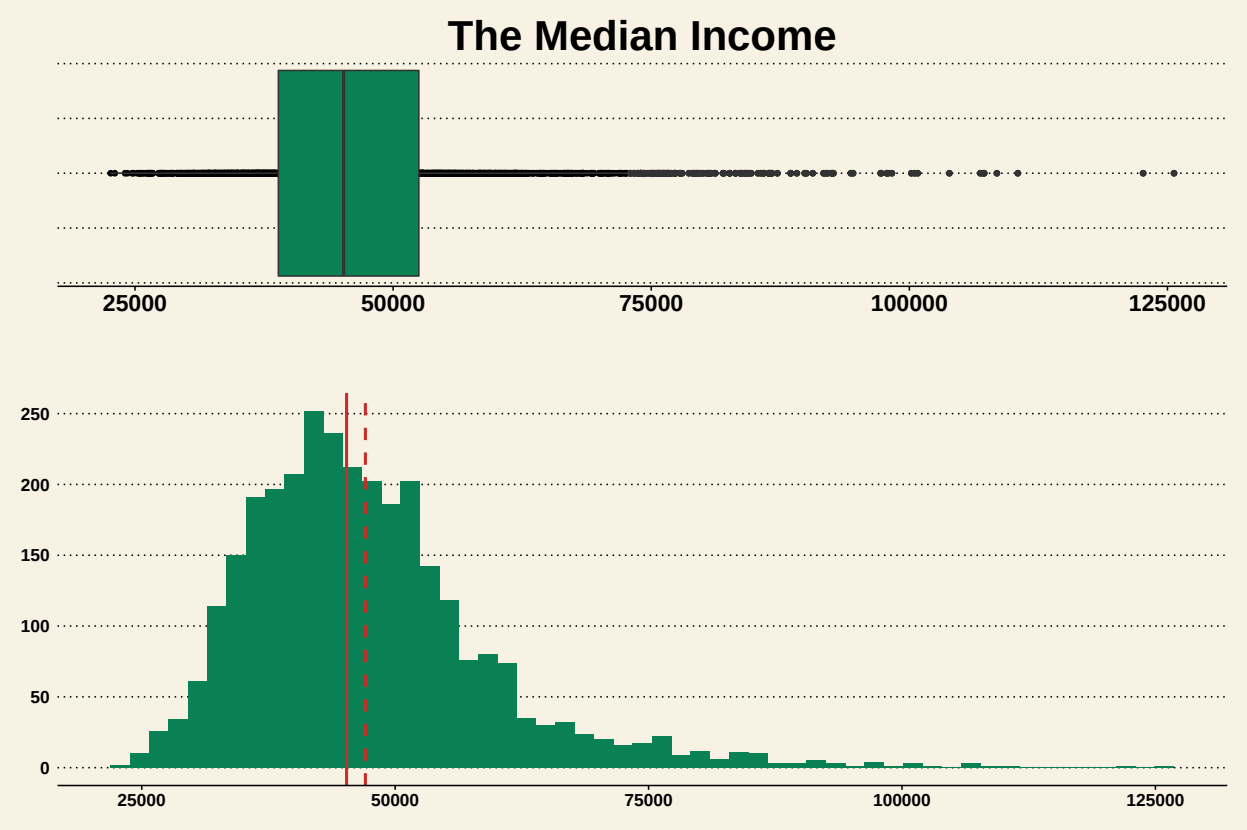
```
summary(Cancer$medIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22640   38883   45207   47063   52492  125635
```

```
sum(Cancer$medIncome > 80000)
```

```
## [1] 64
```

```
boxHist(Cancer$medIncome, "The 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, we could transform it by taking its logarithm. However, we have decided to follow the rule provided by Fox (2011), where logarithmic 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 in two groups: 18-24 and '25 and above' years old. Our initial hypothesis is that the second group should have a stronger correlation with death rate. We validated this hypothesis with the correlations table shown below, that found that only PctBachDeg from the 18-24 group has a correlation with deathRate (although this correlation is very week, -0.31). Instead, as expected, the two'25 and above' education variables have a much higher 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, where we find that these variables indeed seem to be mutually exclusive, given that their range is between 99.9 and 100.1, where the small variations around 100 are likely 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
```

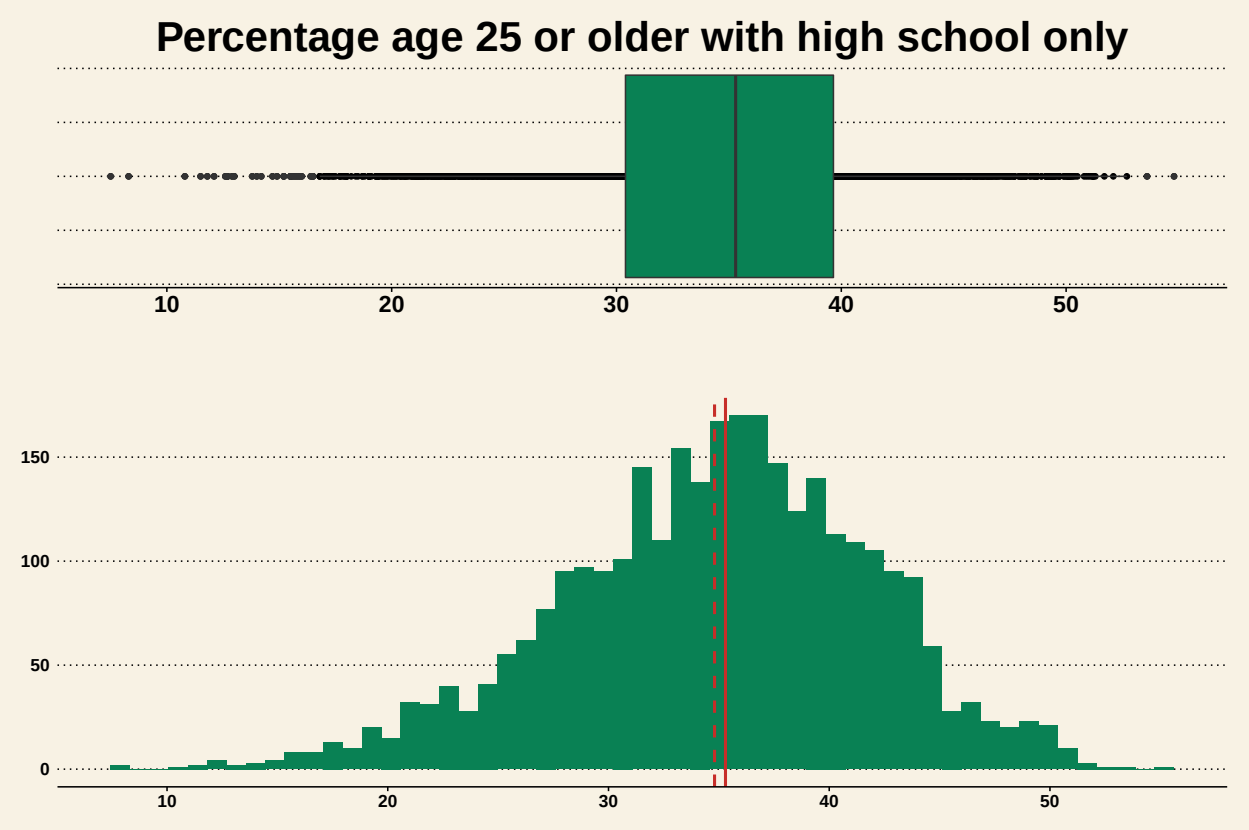
PctHS25_over

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 are no indications 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%)"
```

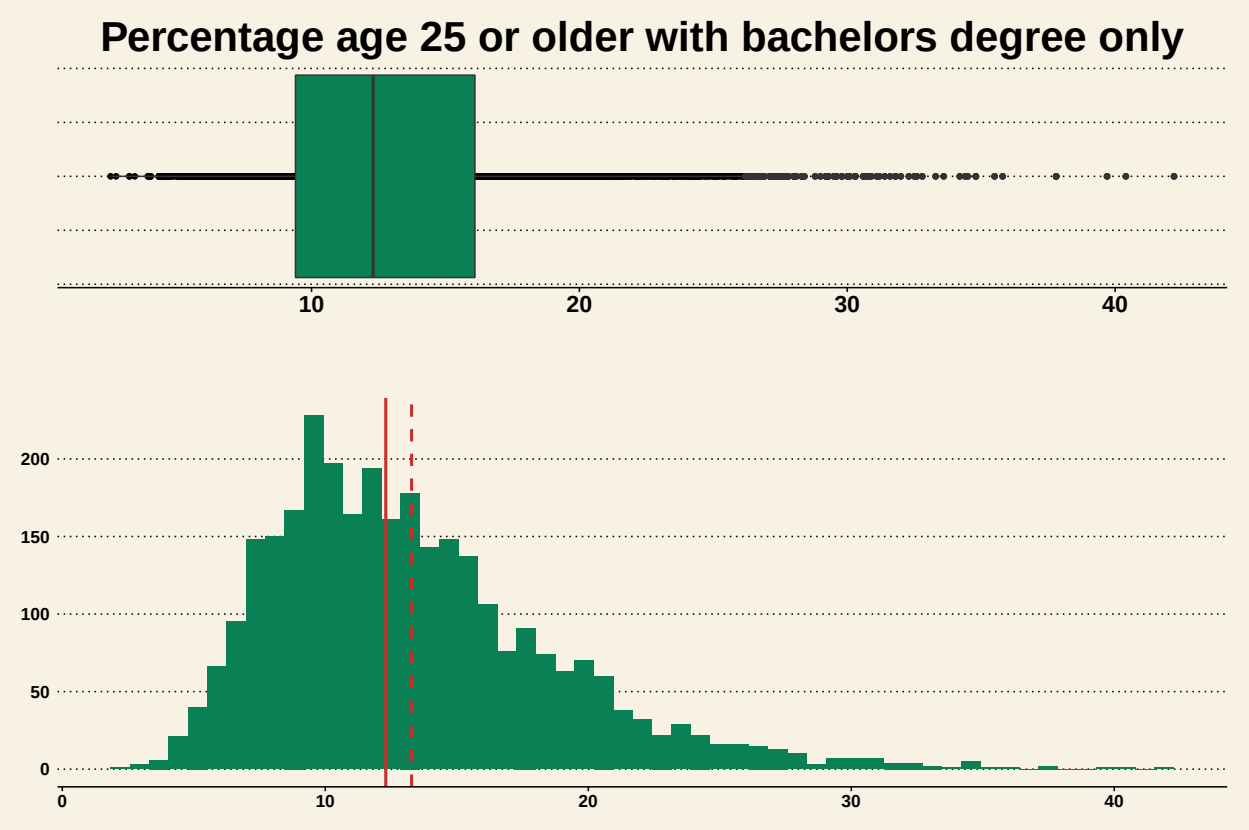
PctBachDeg25_Over

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 will be kept in the data set, 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 Percentage

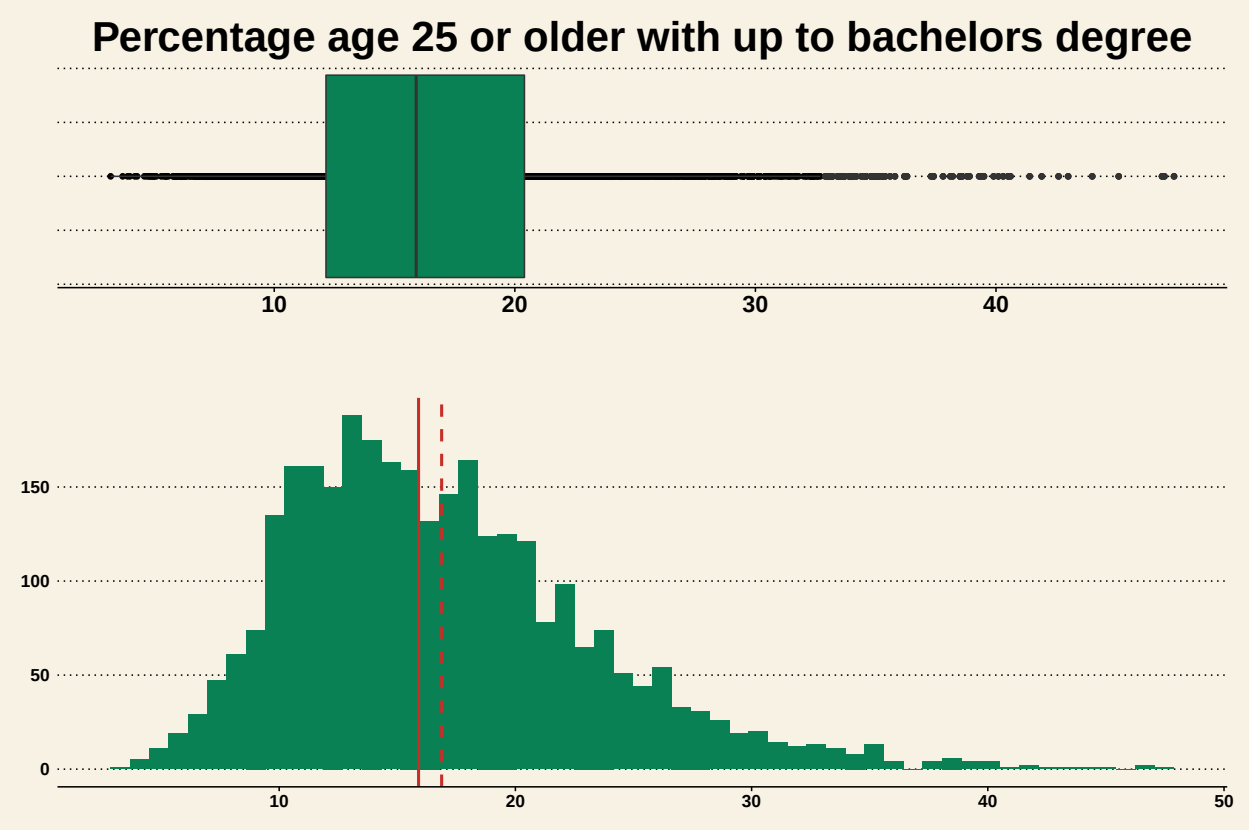
The distribution of `povertyPercent` is unimodal and positively skewed. This is reflected by the fact that all outliers are at the right of the mean. Taking a deeper dive into the outliers, we found that only 3 are extreme while 66 are mild. For this reason, and because we did not find other indication 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 may be necessary to transform it 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%)"
```

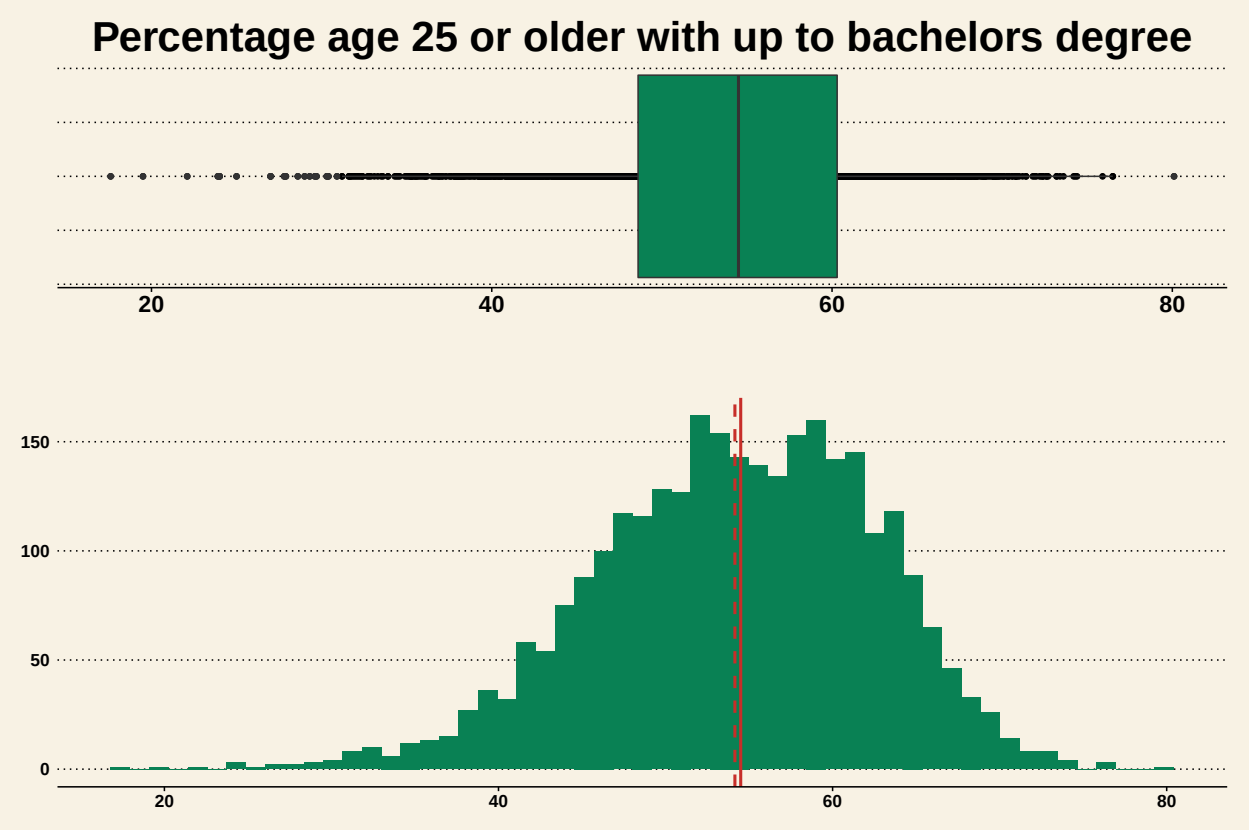
Percentage Employed (16 or over)

The distribution of `PctEmployed16_Over` is unimodal and negatively skewed. This is reflected by the fact that all but one of the outliers are at the left of the mean. There are no extreme outliers and 20 mild outliers (0.7% of observations). 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.

```
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 other 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 between it and `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%)"
```

3. Analysis of Key Relationships

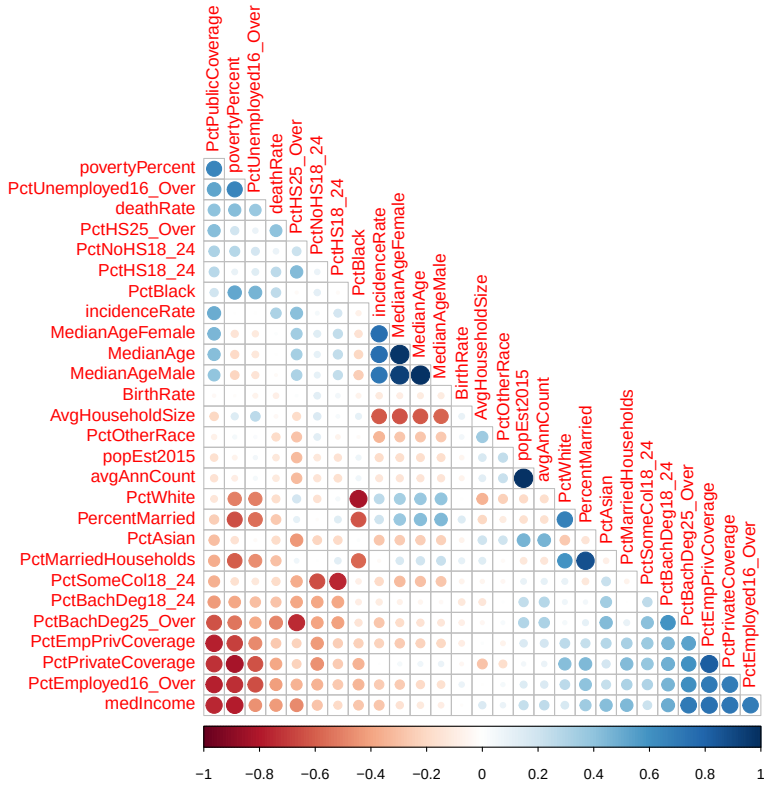
We wanted to find the variables that have strong correlations to the Cancer Incidence (**incedenceRate**), and Cancer Deaths (**deathRate**).

```
Cancer.Numerical   <- Cancer[, !names(Cancer) %in% c('Geography','binnedInc','color','State', 'County')]
Cancer.Correlation <- cor(Cancer.Numerical, use = 'pairwise.complete.obs')
```

We choose all the original numerical variables for this purpose using. We chose to use pairwise complete observations as there were some variables with many missing values that would significantly alter the correlation of other variable pairs otherwise.

Using a matrix correlation plot and using a ordering algorithm to cluster correlated variables together allows us easily find strong POSITIVELY and NEGATIVELY correlated variables.

```
corrplot(Cancer.Correlation, method = 'circle', type = 'lower',  order = 'FPC', diag = F)
```



Network Diagrams

A network diagram allows us to connect the variables (nodes) to each other based on their respective correlation (links). For this we needed to first respahe the Correlation Matrix we generated in the previous step into a flat list of variable combinations along with their respective correlations.

```
links <- subset(melt(Cancer.Correlation), value != 1.0 & abs(value) > 0.4)
links <- links[!duplicated(t(apply(links, 1, sort))),]
```

Afterwards we picked all absolute correlations greater than 0.4, being careful to ignore the diagonal. Also, as the correlation appears twice within the matrix, we removed all duplicate combinations to make the network graph easier to intrepret.

Stylize the links, mapping the inverse log of the correlation magnitudes to the width of the arrows, i.e. stronger the correlation the thicker the arrow. Also, we conditionally mappedchanged the color of the link based on if the correlations are POSITIVE or NEGATIVE.

```
names(links)[1] = 'from'
names(links)[2] = 'to'
names(links)[3] = 'correlation'

links$magnitude <- abs(links$correlation)
links$width     <- 10^links$magnitude
links$color     <- ifelse(links$correlation < 0, 'red', 'blue')
```

The complete set of links were filterd into two subsets that connected to the deathRate or the incidenceRate variable.

```
links.deathRate <- links[links$from == 'deathRate',]
links.incidenceRate <- links[links$from == 'incidenceRate',]
links.outcomeVariables <- rbind(links.deathRate,
                                links.incidenceRate)
```

Next we had to create a list of all nodes–Variables–that we want to see in the diagram. First we start with all the original numerial variables.

```
nodes <- data.frame('id' = names(Cancer.Numerical))
```

Stylize the nodes, highlighting *deathRate* in navy and *incidenceRate* in purple.

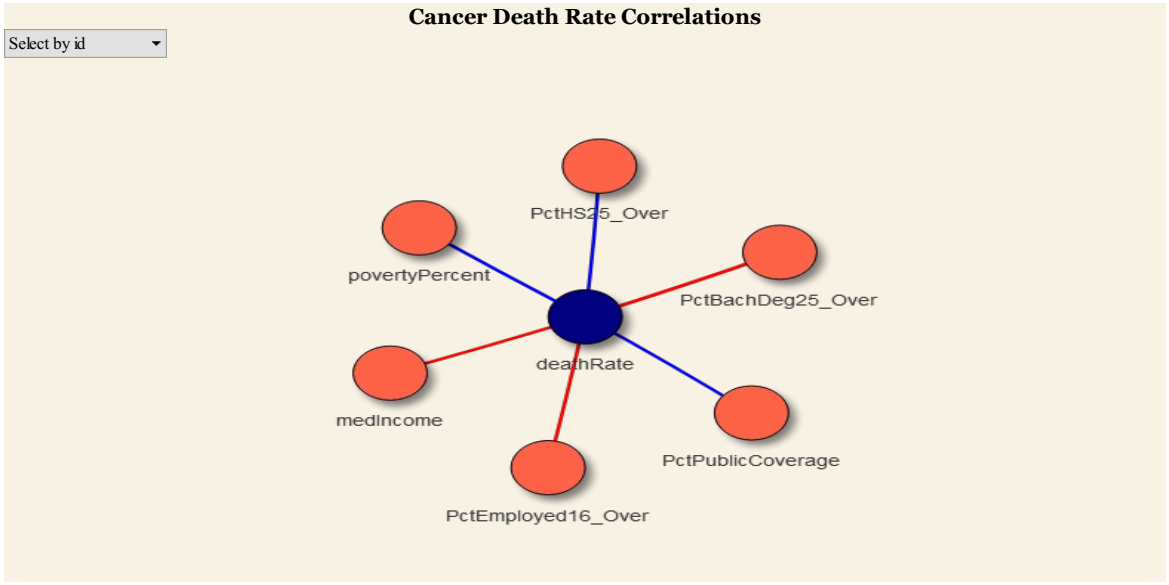
```
nodes$label                      <- nodes$id
nodes$shadow                     <- T
nodes$color.background           <- 'tomato'
nodes$color.border               <- 'black'
nodes$color.highlight.background <- 'orange'
nodes$color.highlight.border     <- 'darkred'

nodes$color.background[nodes$id=='deathRate'    ] = 'navy'
nodes$color.background[nodes$id=='incidenceRate'] = 'purple'

nodes.deathRate <- nodes[nodes$id %in% c(as.vector(links.deathRate$to),'deathRate'),]
nodes.incidenceRate <- nodes[nodes$id %in% c(as.vector(links.incidenceRate$to),'incidenceRate'),]
nodes.outcomeVariables <- nodes[nodes$id %in% unique(c(as.vector(links.incidenceRate$to),
                                                       as.vector(links.deathRate$to),
                                                       'incidenceRate',
                                                       'deathRate')),]
```

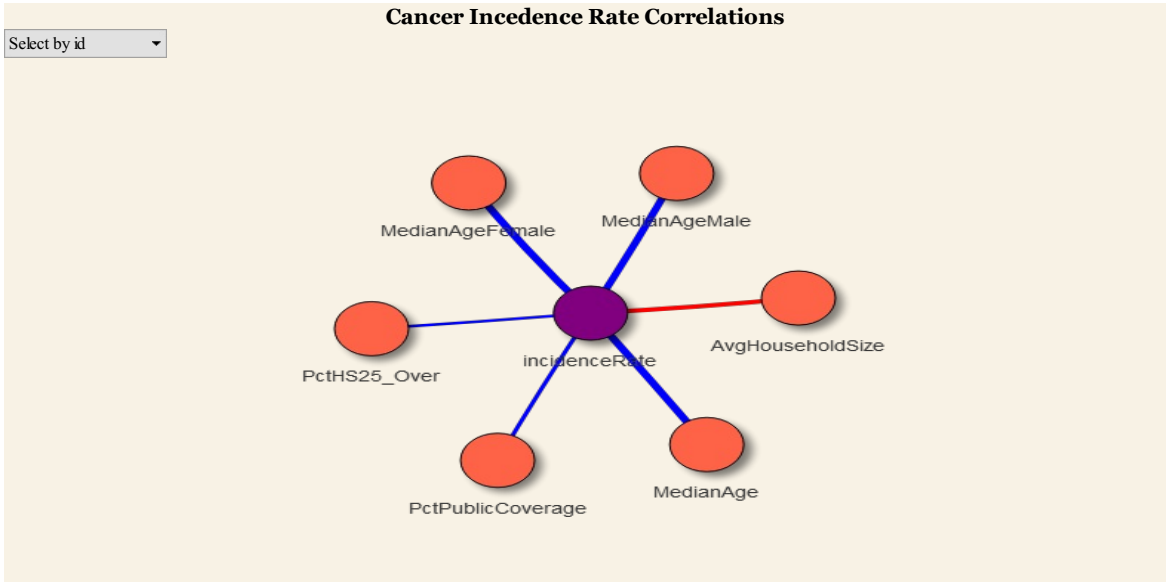
Network diagram showing strong correlations to deathRate.

```
visNetwork(nodes.deathRate, links.deathRate, height = '500px', width = '100%',
           background = '#F8F2E5',
           main = 'Cancer Death Rate Correlations') %>%
    visOptions(highlightNearest = T, nodesIdSelection = T) %>%
    visPhysics(forceAtlas2Based = list(gravitationalConstant = -500))
```



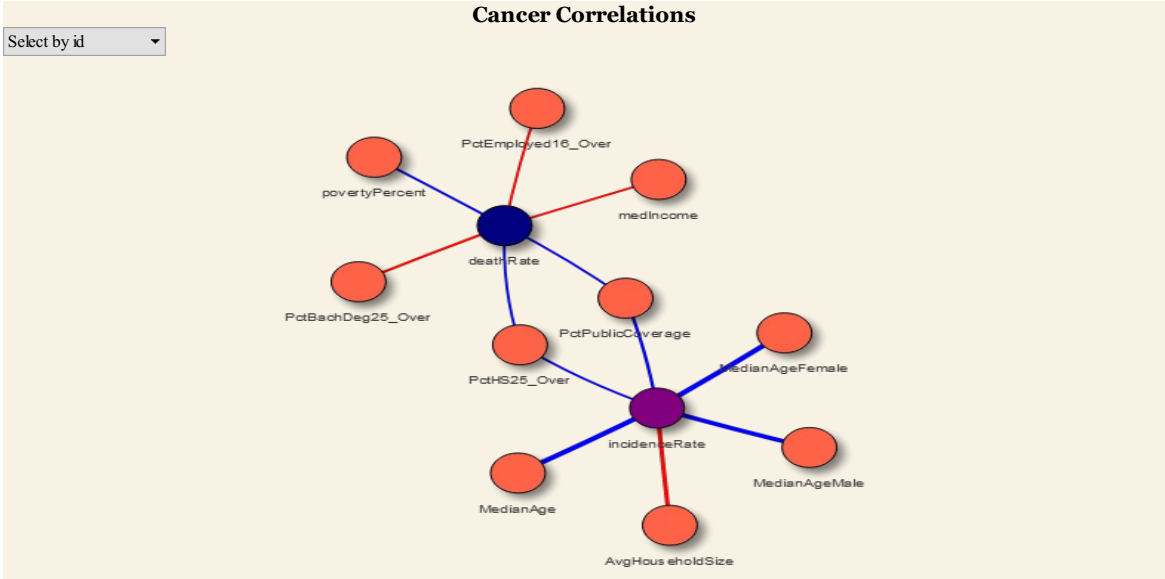
Network diagram showing strong correlations to incidenceRate.

```
visNetwork(nodes.incidenceRate, links.incidenceRate, height = '500px', width = '100%',
           background = '#F8F2E5',
           main = 'Cancer Incedence Rate Correlations') %>%
    visOptions(highlightNearest = T, nodesIdSelection = T)
```



Network diagram showing strong correlations to both outcome variables.

```
visNetwork(nodes.outcomeVariables, links.outcomeVariables, height = '500px', width = '100%',
           background = '#F8F2E5',
           main = 'Cancer Correlations') %>%
    visOptions(highlightNearest = T, nodesIdSelection = T)
```



The preceeding visual analysis helped us to pick the strongest candidates to do further data exploration.

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 be focusing on the former group.

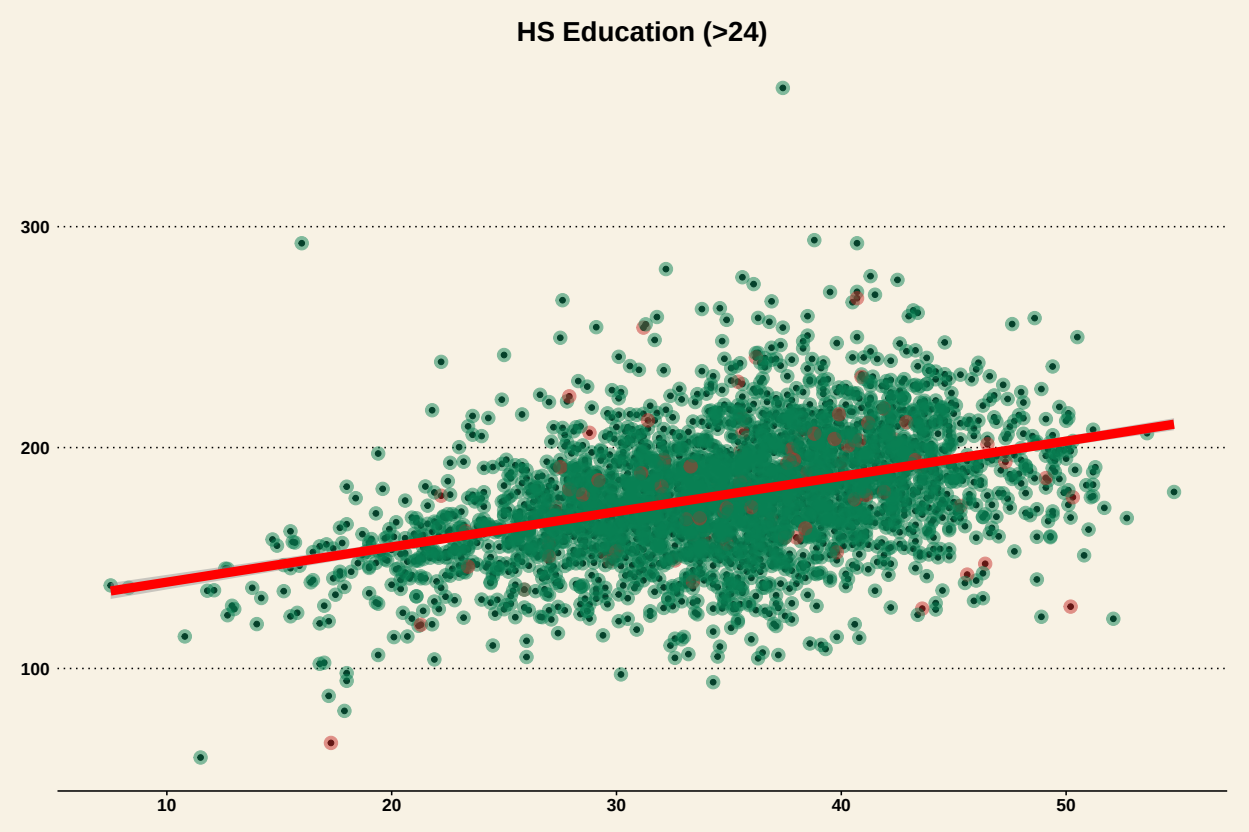
Percentage of Adults with Only a High School Education PctHS25_over

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).

This is an intuitive result since it indicates that a higher concentration of people with low education levels may have poorer health habits and lower access to medical services. However, both of these variables could be affeccted by `MedianAge` in the same direction: older counties might have lower levels of higher education and higher rates of death.

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

```
yxScatter(Cancer$deathRate, Cancer$PctHS25_Over, "HS Education (>24)")
```



Percentage of Adults with atleast a Bachelors Degree [PctBachDeg25_0ver]

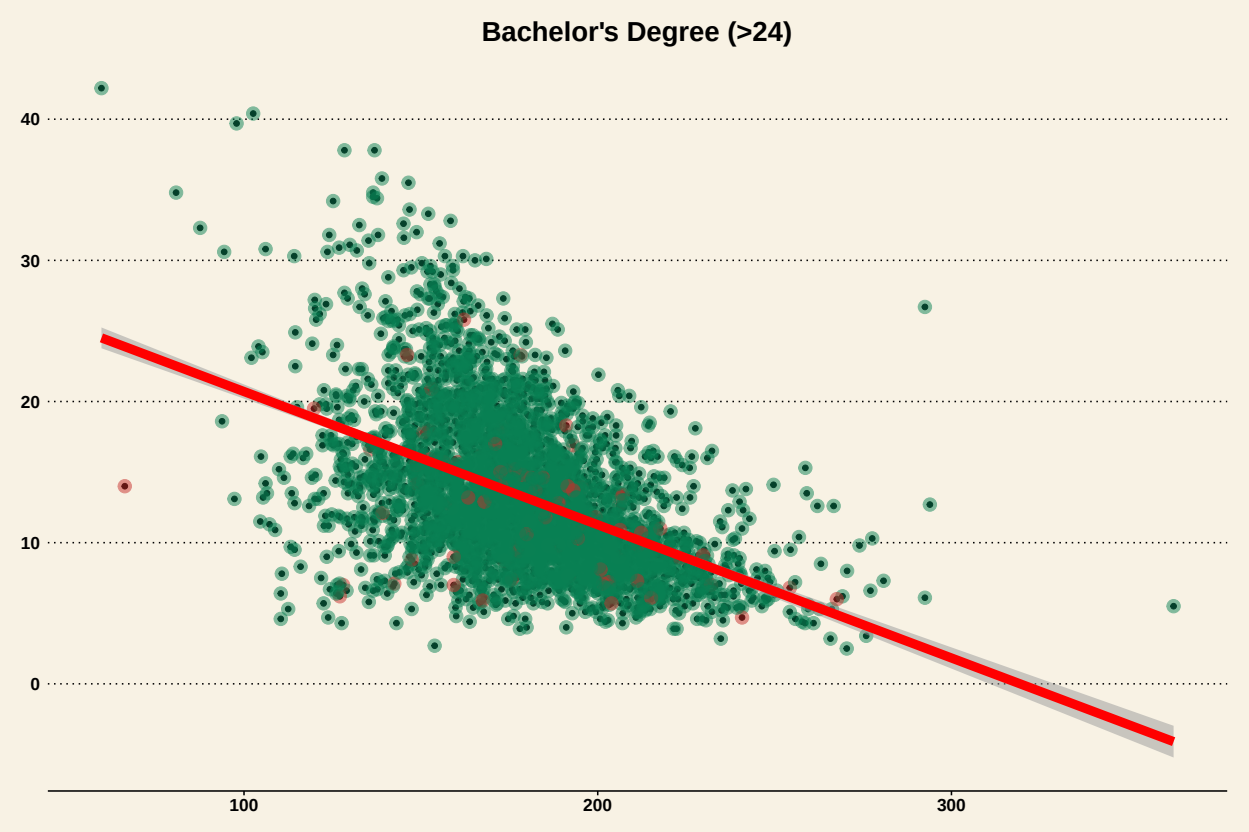
A correlation of -0.48 indicates that there is relationship between `PctBachDeg25_over` and `deathRate`, which is further supported by plotting these variables in a scatterplot, where it can be seen that higher values of percentage of people with bachelors degree are associated to lower levels of death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

This is also an intuitive result, 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.

```
cor(Cancer$deathRate, Cancer$PctBachDeg25_Over)
```

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

```
yxScatter(Cancer$PctBachDeg25_Over, Cancer$deathRate, "Bachelor's Degree (>24)")
```



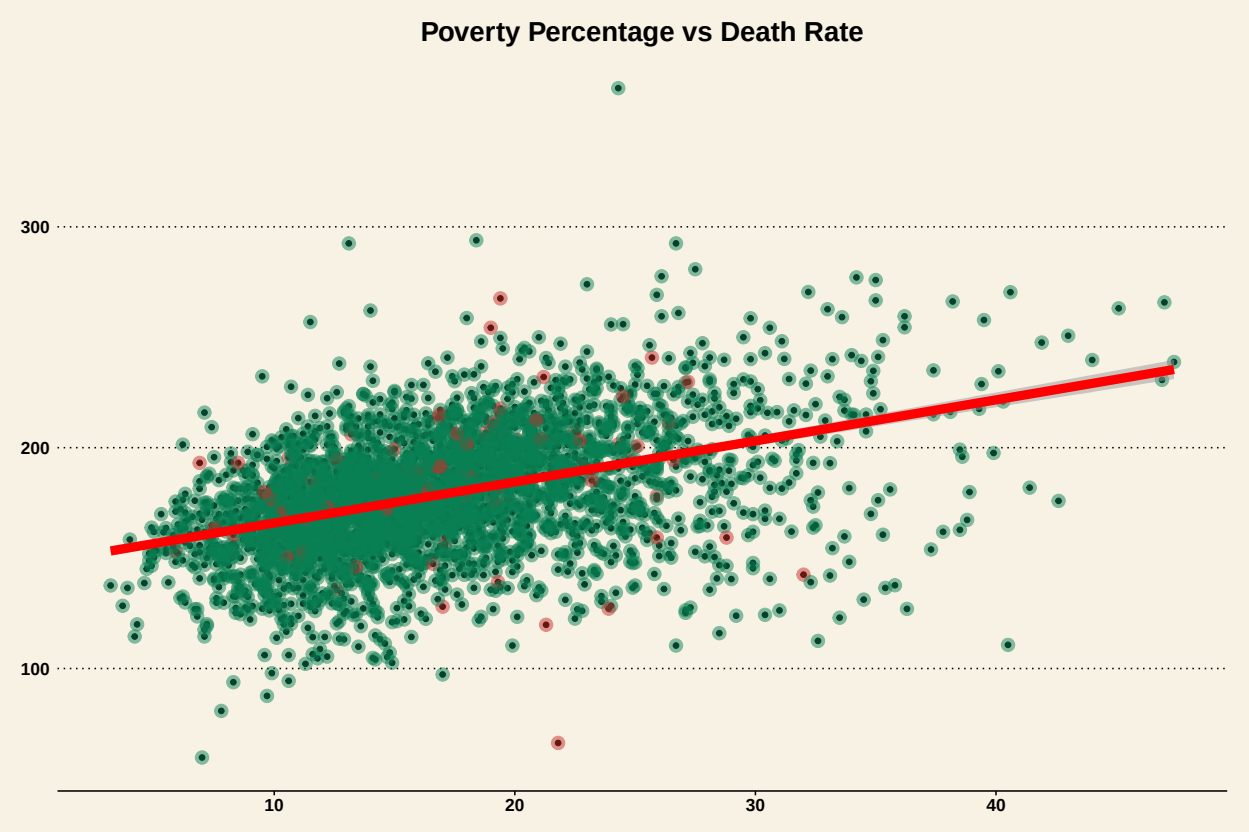
Employment and Poverty

Percentage of Population in Poverty [povertyPercent]

The U.S. government's threshold for poverty in 2015 was $\$24,250$ for a family of four. There is a significant 0.42 correlation between poverty percentage and Cancer death rates.

```
cor(Cancer$deathRate, Cancer$povertyPercent)
## [1] 0.429389
```

```
yxScatter(Cancer$deathRate, Cancer$povertyPercent, "Poverty Percentage vs Death Rate")
```

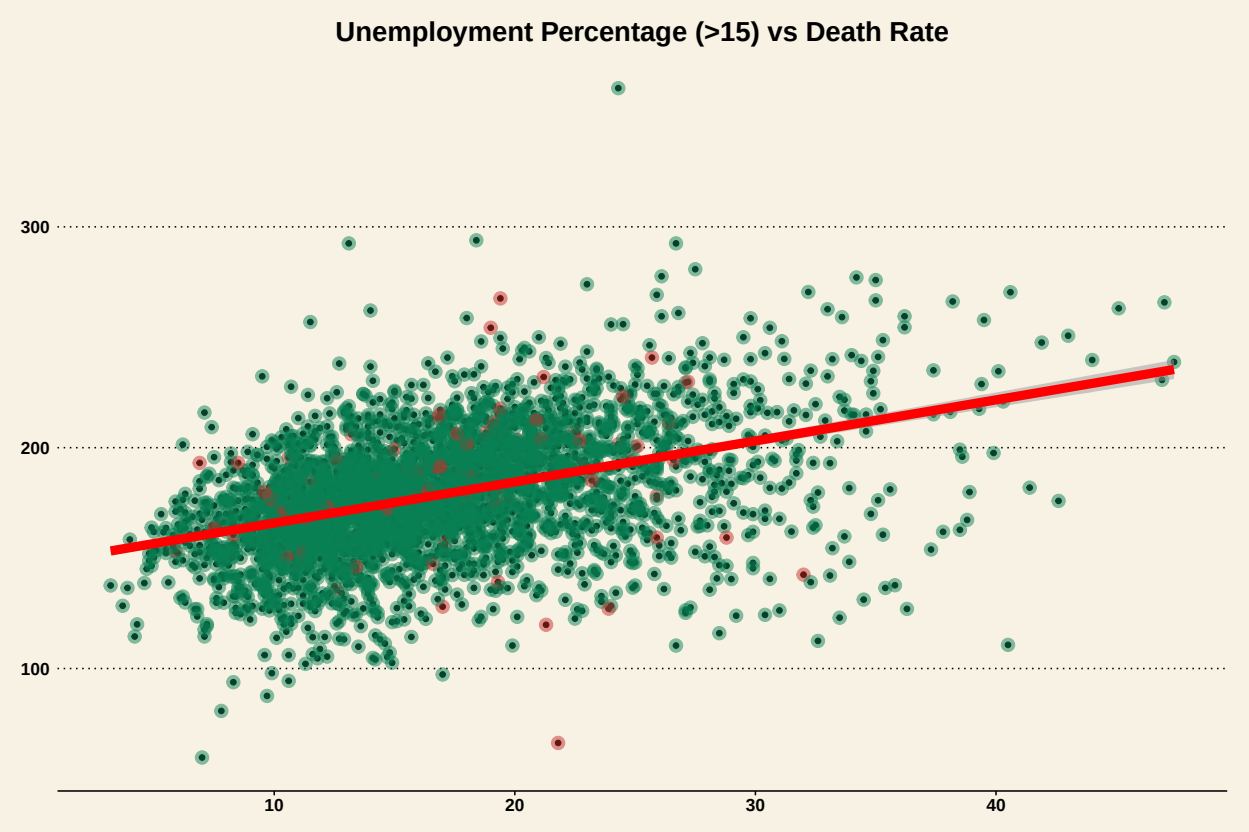


Young Adult and Adult Unemployment Percentage [PctEmployed16_Over]

The percentage of all individuals over 16 years within the county has a significant -0.41 correlation to Cancer death rates. It shows that being employed is a strong correlator to surviving Cancer.

```
cor(Cancer$deathRate, Cancer$PctEmployed16_Over, use = 'complete.obs')
## [1] -0.4120458
```

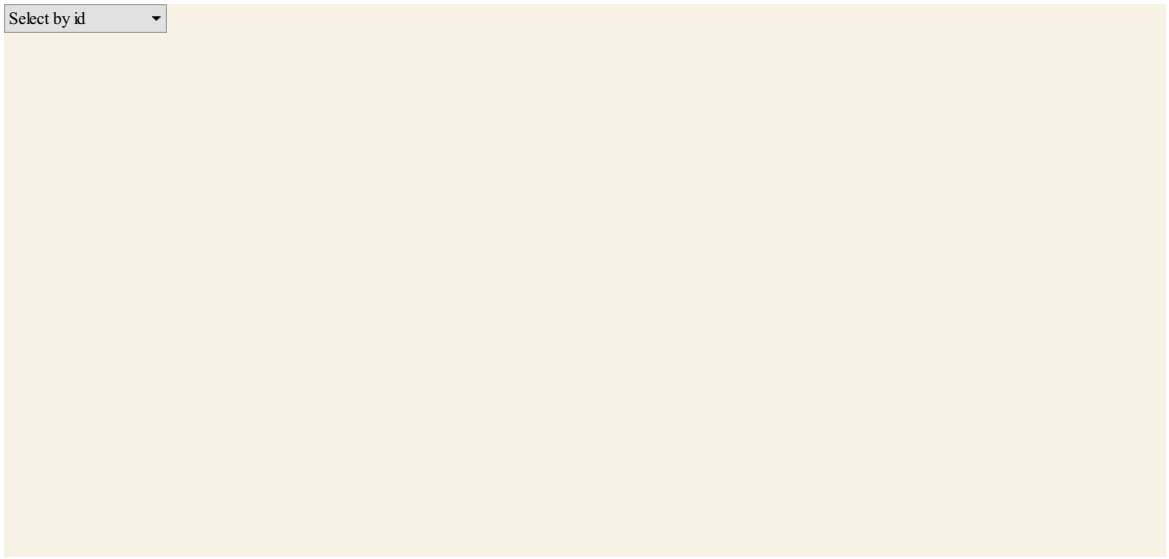
```
yxScatter(Cancer$deathRate, Cancer$povertyPercent, "Unemployment Percentage (>15) vs Death Rate")
```



4. 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 betwee these variables but 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.

```
visNetwork(nodes, links, height = '500px', width = '100%',
           background = '#F8F2E5',
           title = 'Secondary Effects') %>%
    visOptions(highlightNearest = T, nodesIdSelection = T)
```



Age, Family, and Household Size

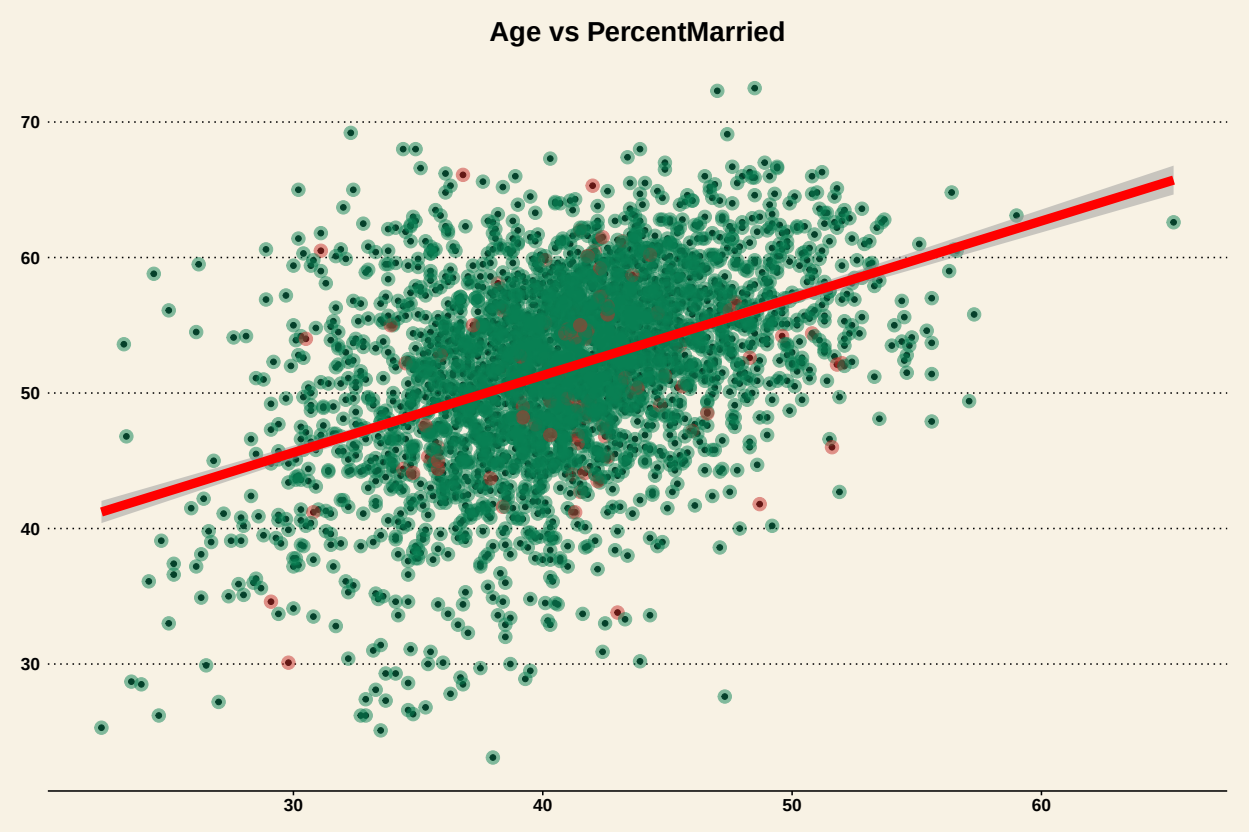
`PercentMarried` has a (weak) relation and `AvgHouseholdSize` has a moderate relation with `MedianAge`. Based on these results, we explored this relationship further.

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

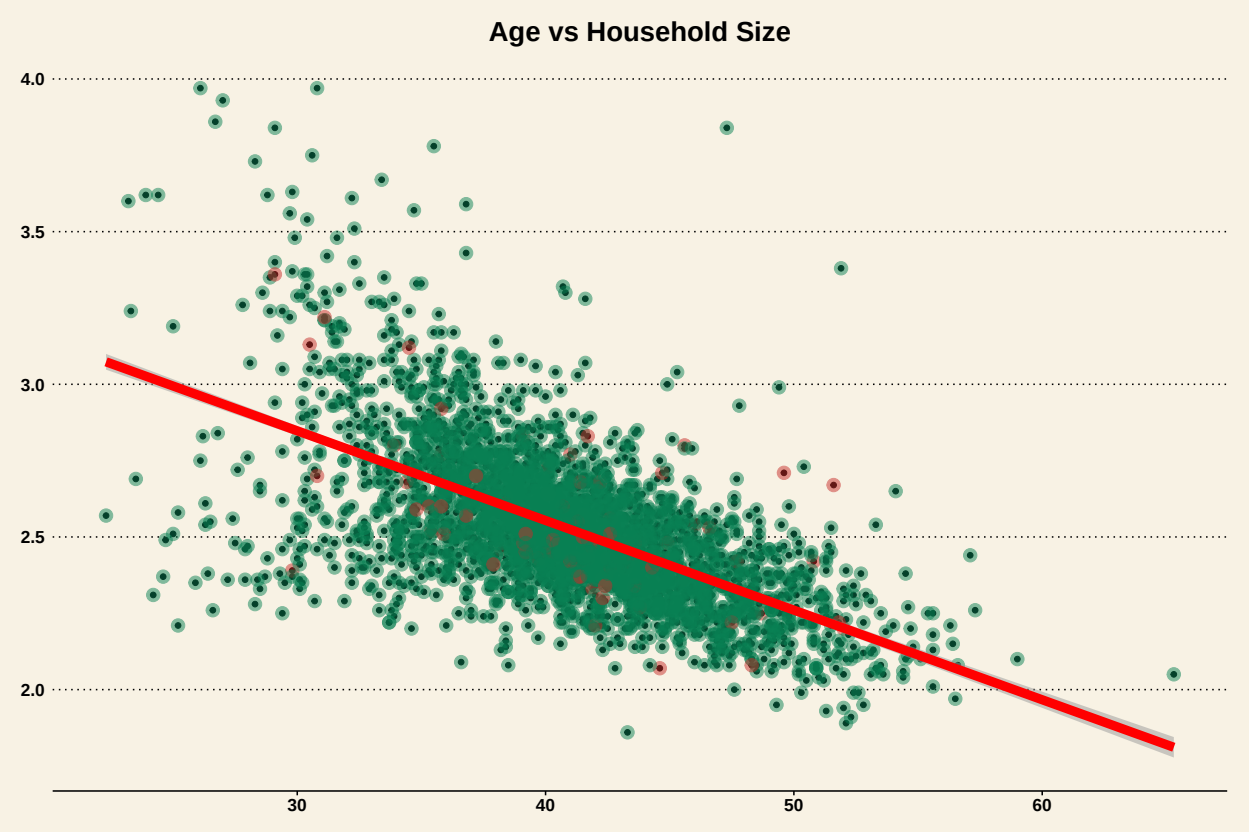
```
##        MedianAge AvgHouseholdSize   PercentMarried
##        1.0000000       -0.6138722        0.4290795
```

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 relation of interest, in order to isolate the effect of the family variables on the death rate.

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




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



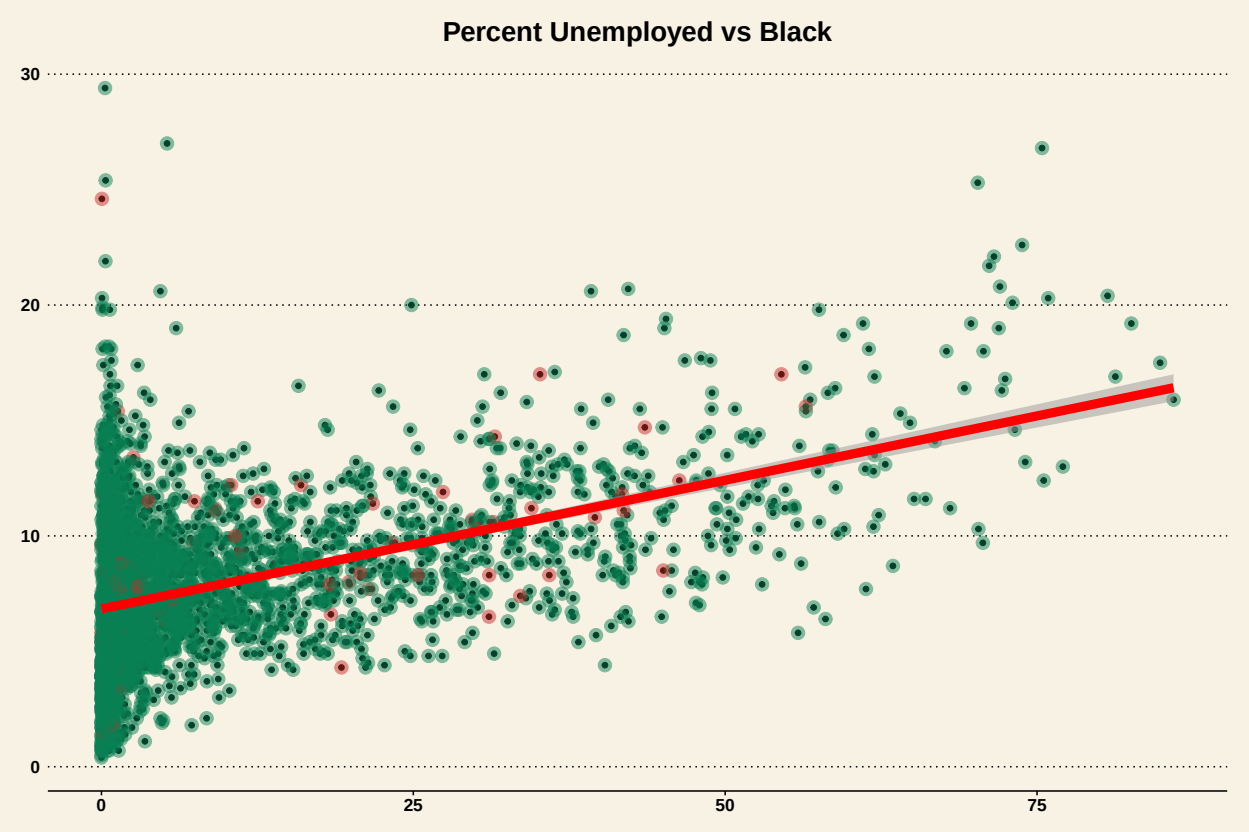
Black Population and Employment

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
```

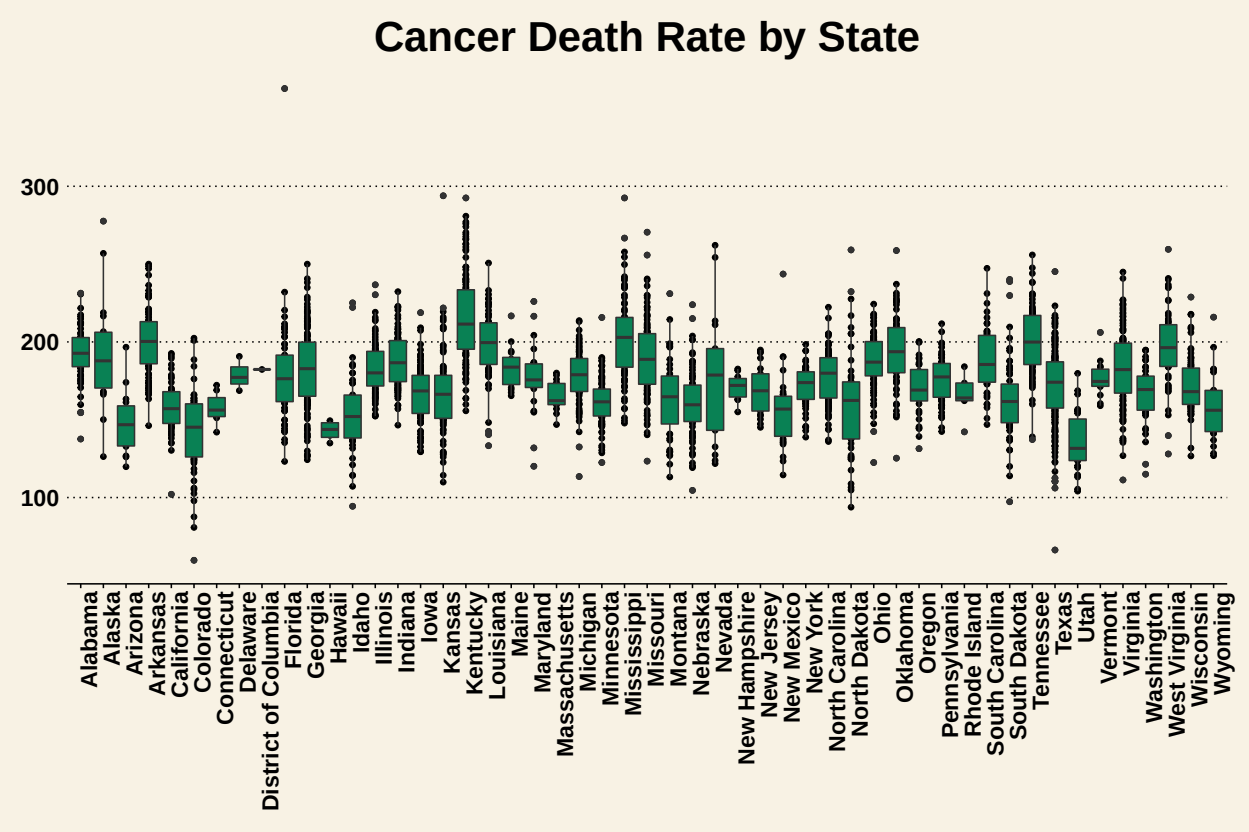
```
#plot(Cancer$PctBlack, Cancer$PctUnemployed16_Over, main = "Age vs PercentMarried")
#abline(lm(PctUnemployed16_Over ~ PctBlack, data = Cancer), lty = 'dashed', lwd = 2, col = 'red')
yxScatter(Cancer$PctUnemployed16_Over, Cancer$PctBlack, "Percent Unemployed vs Black")
```



Death Rates by 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.

```
yxBox(Cancer$deathRate, Cancer$State, 'Cancer Death Rate by State')
```



5. Conclusion