Lab 1 - Ramiro

Ramiro Cadavid

September 23, 2018

Setup

Data transformations

```
Cancer <- read.csv('cancer.csv', row.names = 1)
Cancer <- Cancer %>% separate(Geography, c("County", "State"), sep = ", ", remove = FALSE)
Cancer$MedianAge[Cancer$MedianAge > 100] <- NA
Cancer$AvgHouseholdSize[Cancer$AvgHouseholdSize < 1] <- NA
bins <- seq(20000, 130000, by = 10000)
Cancer$binnedInc2 <- cut(Cancer$medIncome, breaks = bins)
```

2. Univariate Analysis of Key Variables

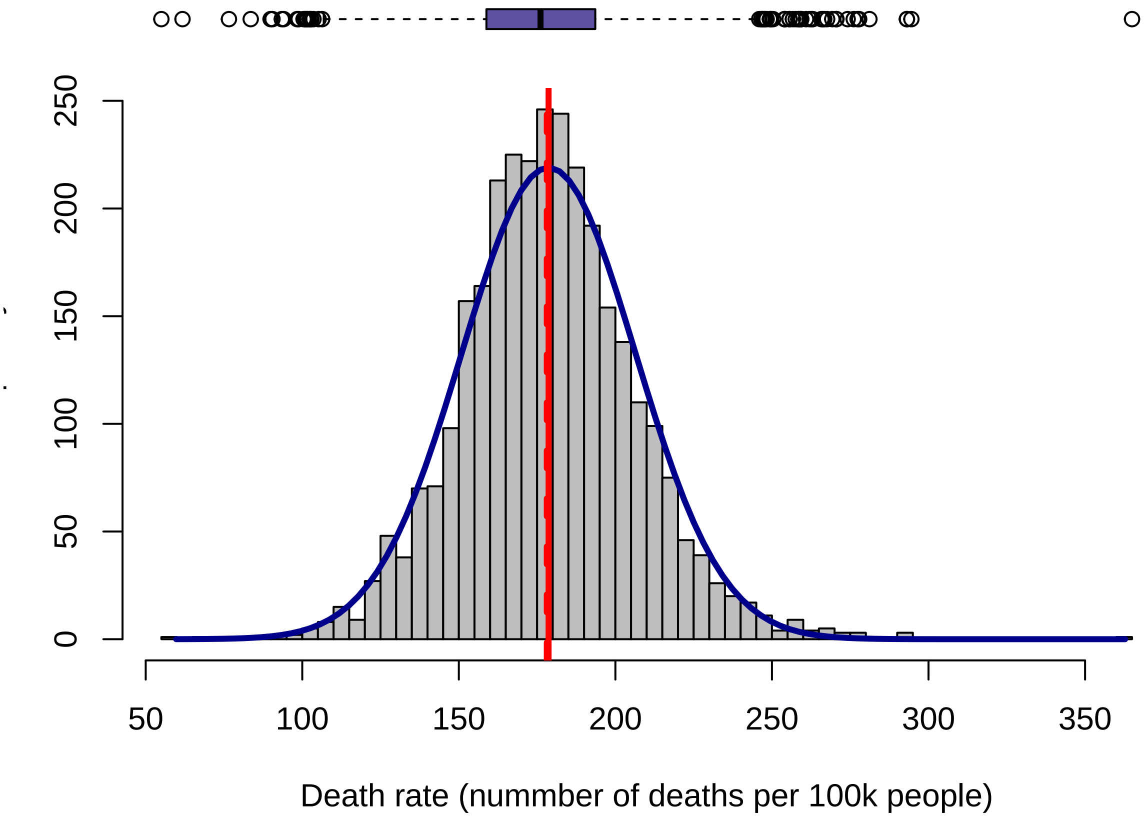
Death rate

- Distribution is symmetric, bell-shaped
- Low proportion of outliers (2.1% of observations)

```
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)")
```



```r
print(paste("Outliers:", length(boxplot.stats(Cancer$deathRate)$out)))
```

```
## [1] "Outliers: 64"
```

```r
print(paste("Outliers (% of total):",
            (length(boxplot.stats(Cancer$deathRate)$out) / nrow(Cancer)) * 100))
```

```
## [1] "Outliers (% of total): 2.10042664916311"
```

Incidence

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

```r
incidence_freq <- data.frame(table(Cancer$avgAnnCount))
incidence_freq[incidence_freq$Freq > 20, ]
```

```
##            Var1 Freq
## 781 1962.667684  206
```

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

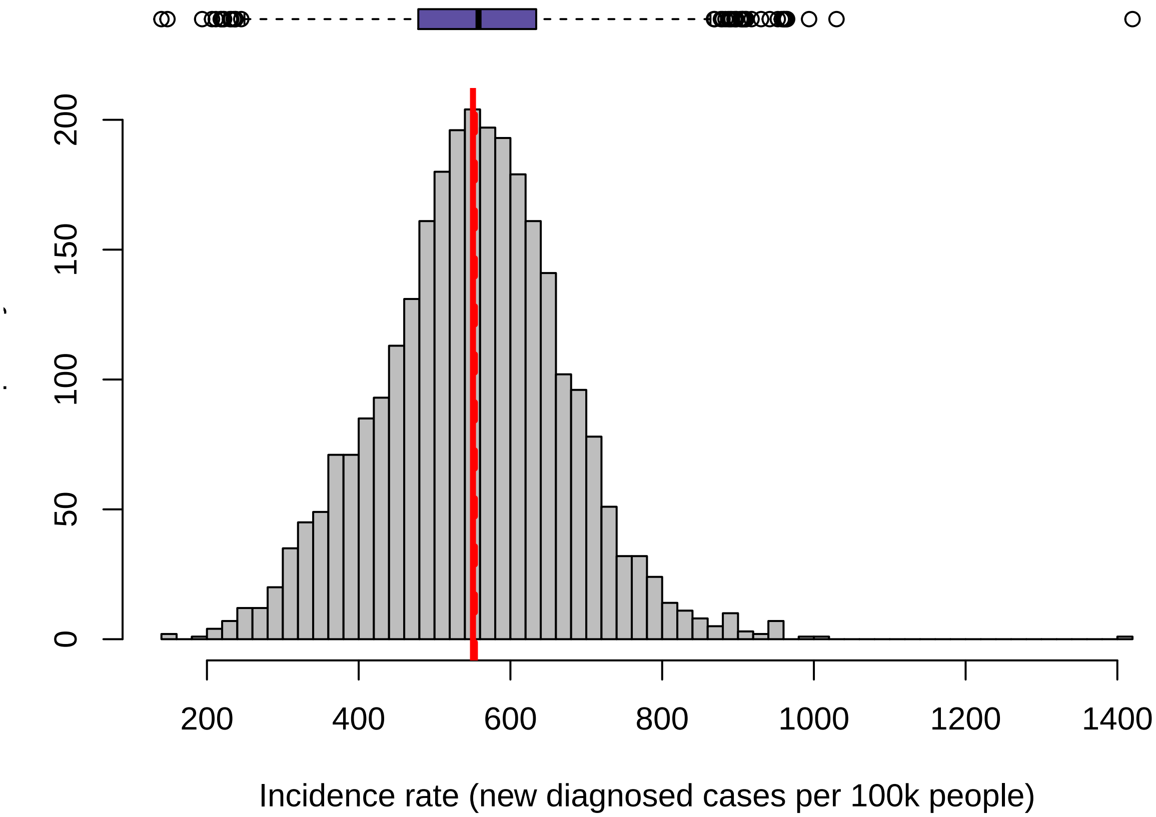
Incidence rate

- The distribution of the incidence rate is unimodal and positively skewed.
- There are 46 outliers. 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, "Incidence rate (new diagnosed cases per 100k people)")
```



```r
print(paste("Outliers:", length(boxplot.stats(Cancer$incidenceRate)$out)))
```

```
## [1] "Outliers: 46"
```

```r
print(paste("Outliers (% of total):",
            (length(boxplot.stats(Cancer$incidenceRate)$out) / nrow(Cancer)) * 100))
```

```
## [1] "Outliers (% of total): 1.50968165408599"
```

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.

```r
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). However, it's distribution is positively sekewed, having 64 counties where the median income is higher than 80,000 USD.

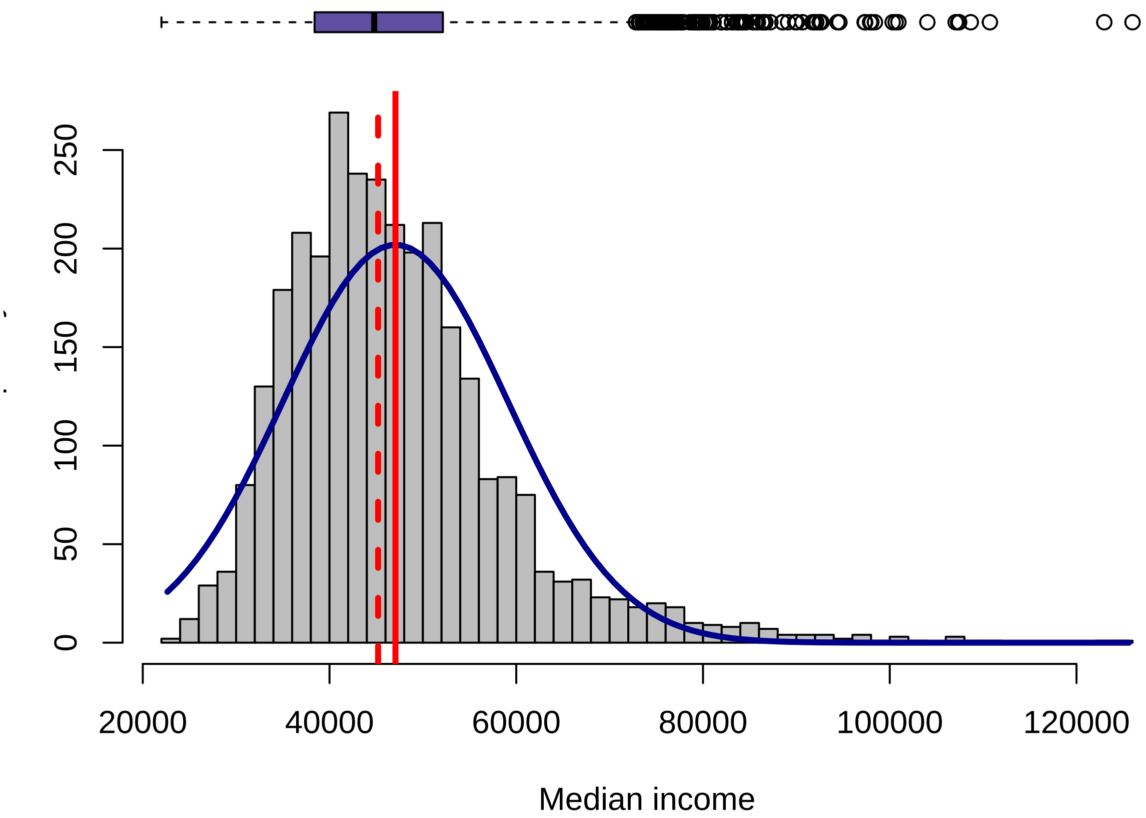
```
summary(Cancer$medIncome)
```

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

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

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

```
boxHist(Cancer$medIncome, "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.

```
print(paste("Outliers:", length(boxplot.stats(Cancer$medIncome)$out)))
```

```
## [1] "Outliers: 122"
```

```
print(paste("Outliers (% of total):", (length(boxplot.stats(Cancer$medIncome)$out) / nrow(Cancer)) * 10
```

```
## [1] "Outliers (% of total): 4.00393829996718"
```

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

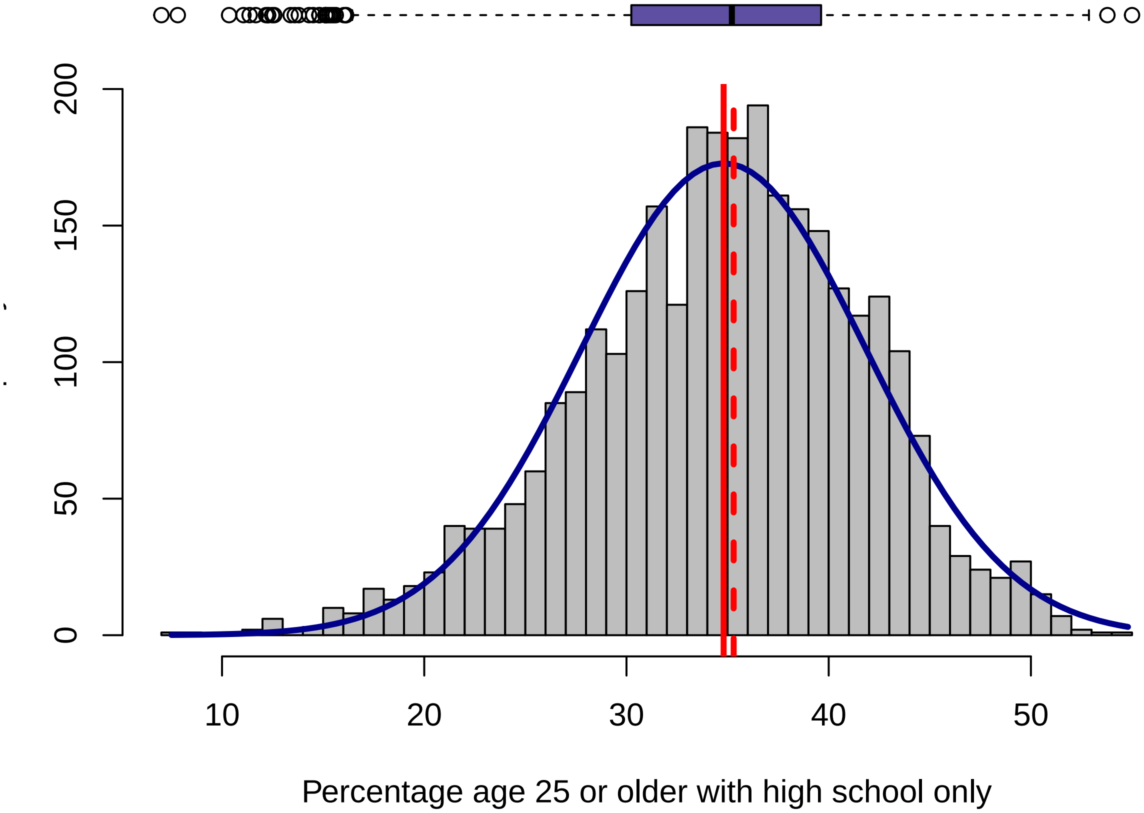
Values in PctHS25 are within a reasonable range (7 to 55%) and there doesn't seem to be an unusual concentration of observations around certain values.

```
summary(Cancer$PctHS25_Over)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.50   30.40   35.30   34.80   39.65   54.80
```

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.

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



```
print(paste("Outliers:", length(boxplot.stats(Cancer$PctHS25_Over)$out)))
```

```
## [1] "Outliers: 31"
```

```
print(paste("Outliers (% of total):", (length(boxplot.stats(Cancer$PctHS25_Over)$out) / nrow(Cancer)) *
```

```
## [1] "Outliers (% of total): 1.01739415818838"
```

Extreme outliers

```
quart.1 <- summary(Cancer$PctHS25_Over)[2]
iqr_hs <- IQR(Cancer$PctHS25_Over)
sum(Cancer$PctHS25_Over < quart.1 - 3*iqr_hs)
```

```
## [1] 0
```

PctBachDeg25_Over

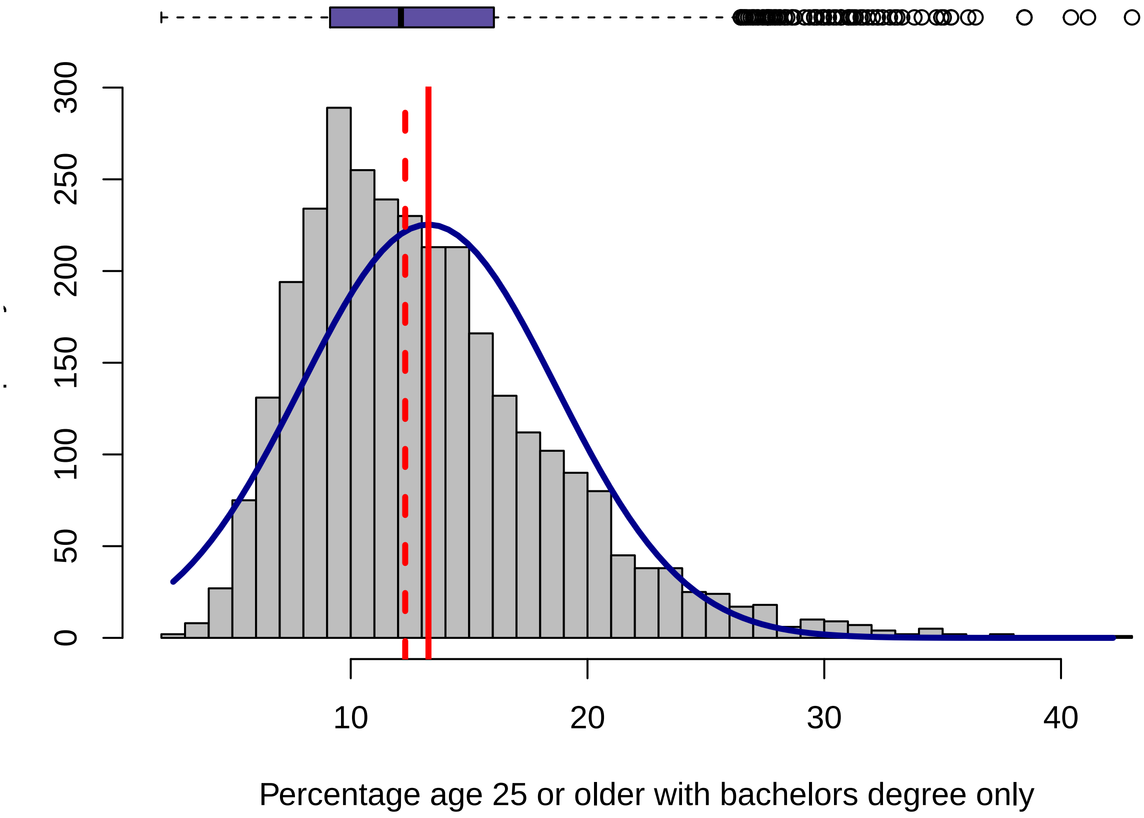
Values in `PctHS25` are within a reasonable range (7 to 55%) and there doesn't seem to be an unusual concentration of observations around certain values.

```
summary(Cancer$PctBachDeg25_Over)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    2.50    9.40   12.30   13.28   16.10   42.20
```

The disribution of this variable is unimodal and positively skewed. It contains 81 outliers (2.7% of observations) all of which at are at the right side of the median. Of these 81 outliers, 31 are extreme outliers, that need to be taken into account when modeling the relationship between this variable and `deathRate` but need to be kept in the data set, since there are no indications that they are errors.

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



```
print(paste("Outliers:", length(boxplot.stats(Cancer$PctBachDeg25_Over)$out)))
```

```
## [1] "Outliers: 82"
```

```
print(paste("Outliers (% of total):", (length(boxplot.stats(Cancer$PctBachDeg25_Over)$out) / nrow(Cance
```

```
## [1] "Outliers (% of total): 2.69117164424024"
```

```
summary(boxplot.stats(Cancer$PctBachDeg25_Over)$out)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.20   27.30   29.30   30.08   31.75   42.20
```

Extreme outliers

```
quart.1 <- summary(Cancer$PctBachDeg25_Over)[2]
iqr_hs <- IQR(Cancer$PctBachDeg25_Over)
sum(Cancer$PctBachDeg25_Over > quart.1 + 3*iqr_hs)
```

```
## [1] 39
```

3. 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 be focusing on the former group.

PctHS25_over

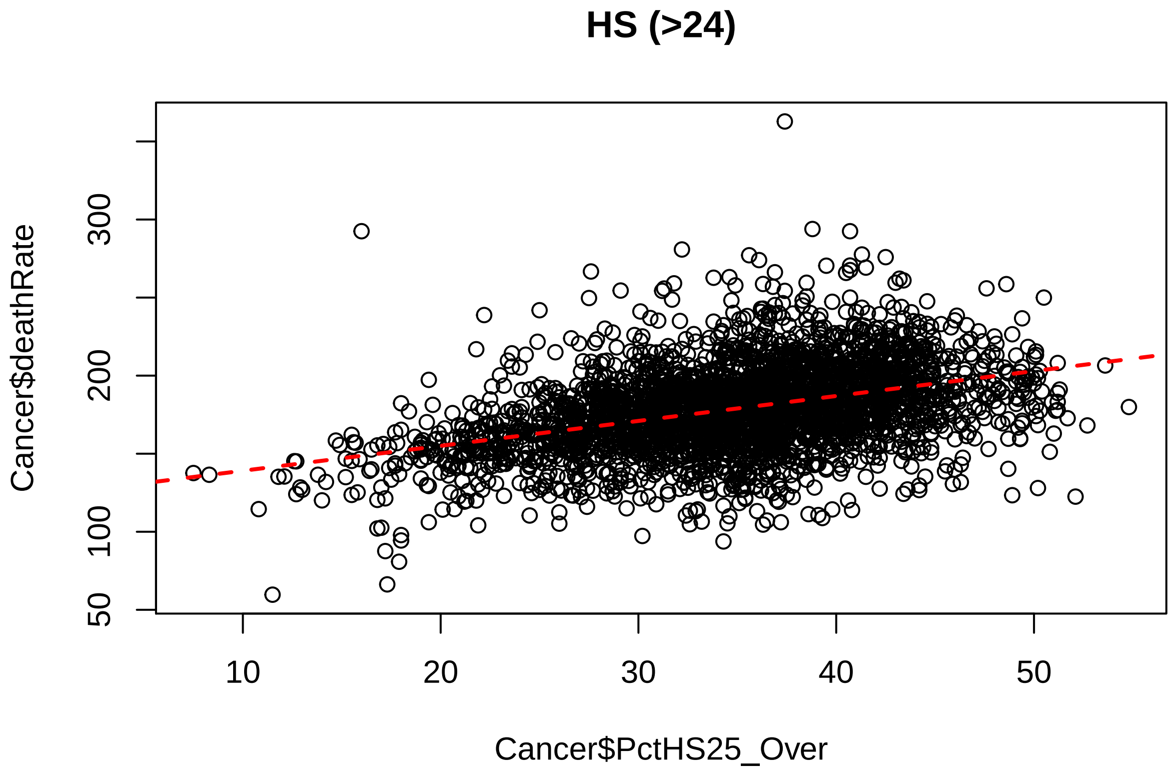
A correlation of 0.4 between `PctHS25_over` and `deathRate` indicates that there is indeed a relationship between these variables, which is further indicated by plotting them together in a scatterplot, that shows that higher values of percentage of population with only high school tend to be associated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

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; then, it is necessary to test this relationship in the following section, to find if age confounds this relationship.

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

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

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



PctBachDeg25_Over

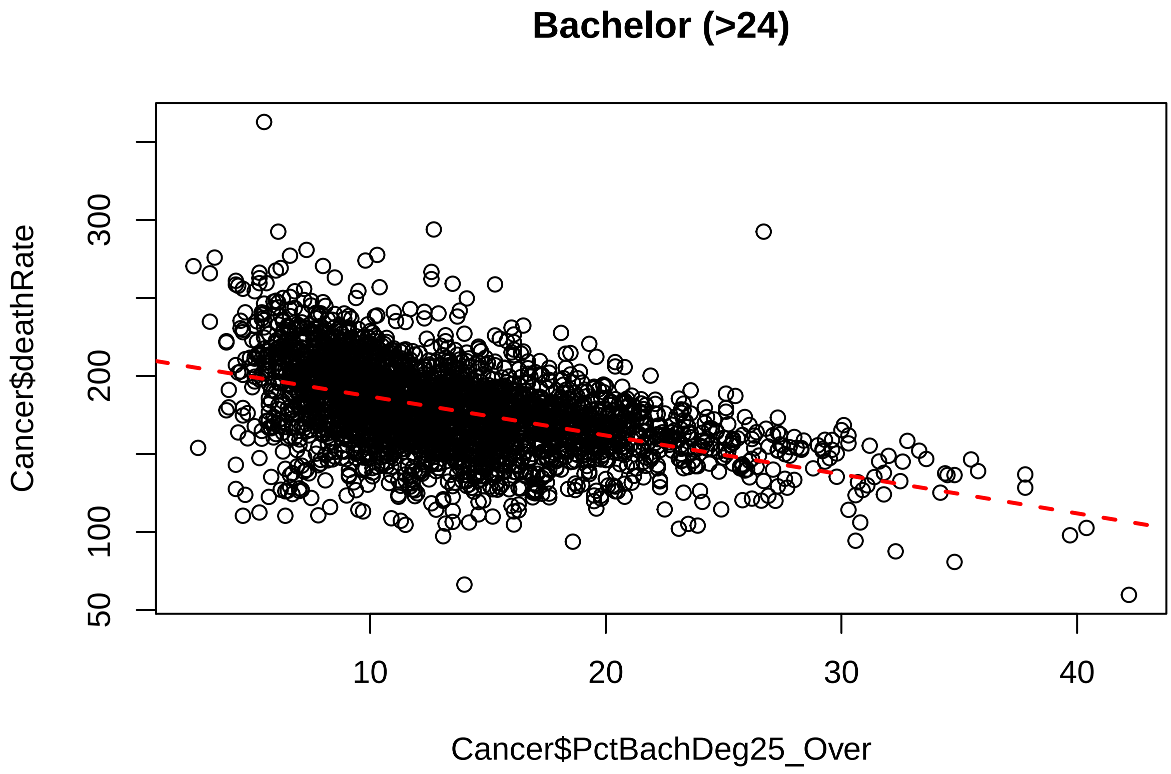
A correlation of −0.48 indicates that there is relationship between `PctBachDeg25_over` and `deathRate`, which is further supported by plotting these variables in a scatterplot, where it can be seen that higher values of percentage of people with bachelors degree are associated to lower levels of death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

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



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 these relationships but may be confounded by one or more variables. To further assess this systematically, the following network visualization shows the variables that have a correlation higher than xxxxxxxx, where each node represents a different variable and each vertex indicates the strength of the relationship between the variables connected.

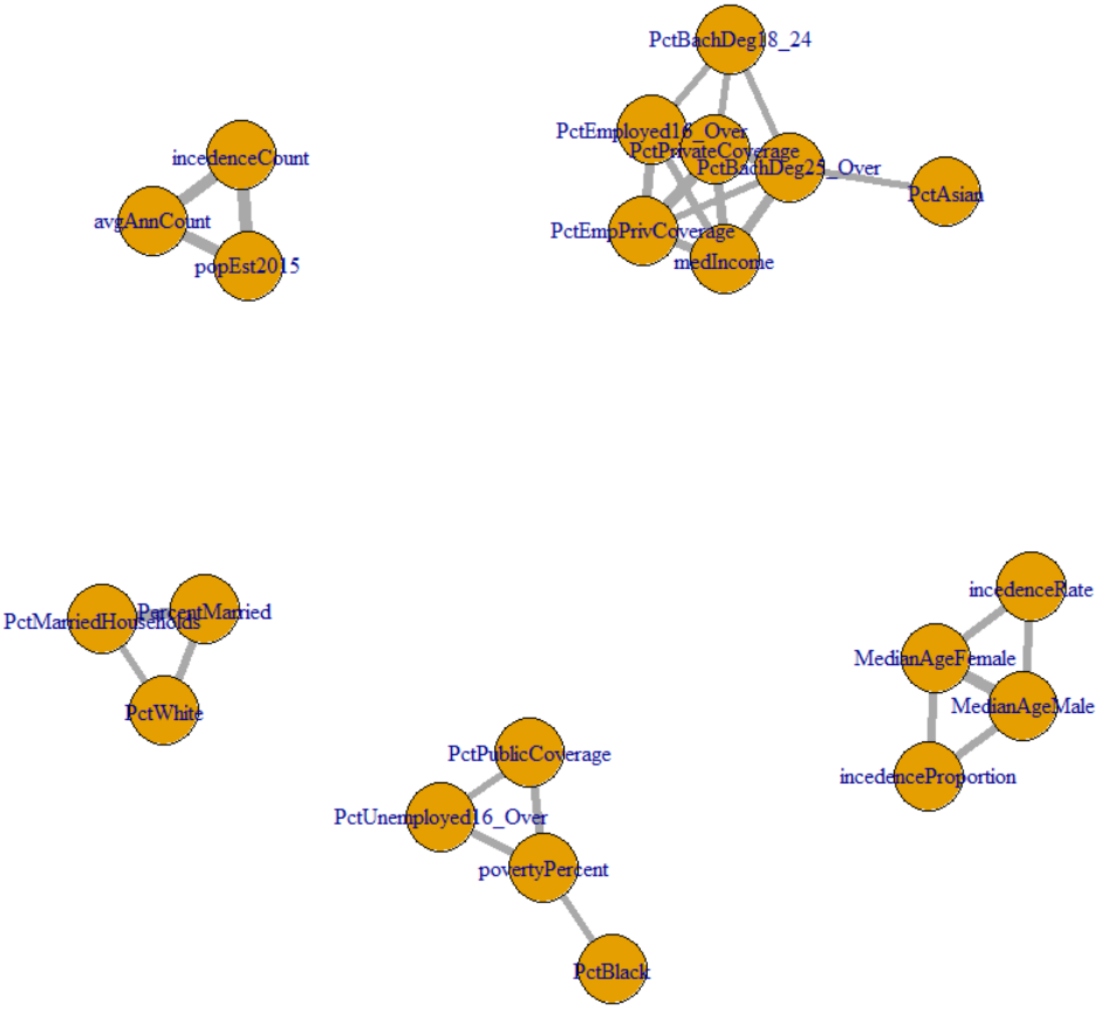


Figure 1: secondary analysis

Based on this network,