Cancer EDA_Tim_Mike_Craig_W203_4

Tim, Mike, Craig, Wei

1/21/2018

Introduction

This report presents an initial exporatory data analysis identifying the key features associated with cancer rates and deaths based on geographic location in the form of a county in the United States. The goal is to use the findings from the analysis to develop strategies to improve future cancer outcomes. We are a team of data scientists motivated to in promote understanding the societal factors that impact mortality rates of cancer among various communities in the United States. We are grateful for the grant awarded to us by a health government agency to complete this study.

Research Question

Our task was to answer these two key research quesitons:

1. What are the key county level characteristics associated with mortality rates from cancer?
2. Are there trends in county level characteristics that can be identified to inform social intervention to decrease cancer mortality rates?

About the Data

```
df <- read.csv("cancer.csv")
class(df)  # dimensions are 3047 rows x 30 columns
## [1] "data.frame"
dim(df)  # dimensions of the data set
## [1] 3047   30
```

The dataset that we explored had 3047 observations and 30 variables. The majority of the data types are either number or integer with Geography and binnedInc as factors. The death rate was designated the target variable by the health government agency. Each observation represents a single county within the United States, and each variable describes that county for a number of different ways. The dataset provided came with a minimal data dictionary and is presented in Table 1. Most of the variables not defined were self explanatory by key notations in the variable name such 'Pct' for percent, 'Avg' for average, and 'med' for median. The units for death rate and birth rates were not provided and without a means to confirm the units we chose to leave these variables without units. We further grouped variables into categories based on similiarities to help organize the analysis. Some data processing was required to address select issues as described in the section *Data Quality*.

Table 1. Data Dictionary:

Variable Name	Variable Description	category
deathRate*	rate of deaths due to cancer	Outcome
avgAnnCount	2009 - 2013 mean incidents of cancer per county.	Cancer incidence
Geography*	Description of county location (county name, state)	County
popEst2015	County population, estimated during 2015	Population/Birth
BirthRate*		Population/Birth

Variable Name	Variable Description	category
MedianAge*	Median age of county residents	Age
MedianAgeMale*	Median age of Male county residents	Age
MedianAgeFemale*	Median age of Female county residents	Age
PctWhite*	% of residents with race designation: White	Ethnicity
PctBlack*	% of residents with race designation: Black	Ethnicity
PctAsian*	% of residents with race designation: Asian	Ethnicity
PctOtherRace*	% of residents with race designation: Other	Ethnicity
AvgHouseholdSize*	Average household size, number of people	Household/Marital
PercentMarried*	% maried	Household/Marital
PctMarriedHouseholds*	% of households that are married	Household/Marital
PctNoHS18_24*	% of residents, age 18-24, without completing a high school education	Education
PctHS18_24*	% of residents, age 18-24, with a high school degree (highest education)	Education
PctSomeCol18_24*	% of residents, age 18-24, with some college education (no degree)	Education
PctBachDeg18_24*	% of residents, age 18-24, with a bachelor's degree	Education
PctHS25_Over*	% of residents, age over 25, with a high school degree (highest education)	Education
PctBachDeg25_Over*	% of residents, age over 25, with a bachelors degree	Education
PctPrivateCoverage	% of residents with private insurance coverage	Insurance
PctEmpPrivCoverage*	% of residents with employee-provided private insurance coverage	Insurance
PctPublicCoverage	% of residents with public insurance coverage	Insurance
PctEmployed16_Over*	% of residents, age over 16, that are employed	Employment
PctUnemployed16_Over*	% of residents, age over 16, that are unemployed	Employment
medIncome*	Median Income	Income
povertyPercent	% of county residents living below the poverty line	Income
binnedInc*	Binned income level	Income

```r
str(df)  #names of columns and variable types
```

```
## 'data.frame':    3047 obs. of  30 variables:
##  $ X                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ avgAnnCount        : num  1397 173 102 427 57 ...
##  $ medIncome          : int  61898 48127 49348 44243 49955 52313 37782 40189 42579 60397 ...
##  $ popEst2015         : int  260131 43269 21026 75882 10321 61023 41516 20848 13088 843954 ...
##  $ povertyPercent     : num  11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
##  $ binnedInc          : Factor w/ 10 levels "(34218.1, 37413.8]",..: 9 6 6 4 6 7 2 2 3 8 ...
##  $ MedianAge          : num  39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
##  $ MedianAgeMale      : num  36.9 32.2 44 42.2 47.8 43.5 42.2 50.8 48.4 34.7 ...
##  $ MedianAgeFemale    : num  41.7 33.7 45.8 43.4 48.9 48 43.5 52.5 49.8 37 ...
##  $ Geography          : Factor w/ 3047 levels "Abbeville County, South Carolina",..: 1459 1460 1464
##  $ AvgHouseholdSize   : num  2.54 2.34 2.62 2.52 2.34 2.58 2.42 2.24 2.38 2.65 ...
##  $ PercentMarried     : num  52.5 44.5 54.2 52.7 57.8 50.4 54.1 52.7 55.9 50 ...
##  $ PctNoHS18_24       : num  11.5 6.1 24 20.2 14.9 29.9 26.1 27.3 34.7 15.6 ...
##  $ PctHS18_24         : num  39.5 22.4 36.6 41.2 43 35.1 41.4 33.9 39.4 36.3 ...
##  $ PctSomeCol18_24    : num  42.1 64 NA 36.1 40 NA NA 36.5 NA NA ...
##  $ PctBachDeg18_24    : num  6.9 7.5 9.5 2.5 2 4.5 5.8 2.2 1.4 7.1 ...
##  $ PctHS25_Over       : num  23.2 26 29 31.6 33.4 30.4 29.8 31.6 32.2 28.8 ...
##  $ PctBachDeg25_Over  : num  19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
##  $ PctEmployed16_Over : num  51.9 55.9 45.9 48.3 48.2 44.1 51.8 40.9 39.5 56.6 ...
```

```
##  $ PctUnemployed16_Over: num  8 7.8 7 12.1 4.8 12.9 8.9 8.9 10.3 9.2 ...
##  $ PctPrivateCoverage  : num  75.1 70.2 63.7 58.4 61.6 60 49.5 55.8 55.5 69.9 ...
##  $ PctEmpPrivCoverage  : num  41.6 43.6 34.9 35 35.1 32.6 28.3 25.9 29.9 44.4 ...
##  $ PctPublicCoverage   : num  32.9 31.1 42.1 45.3 44 43.2 46.4 50.9 48.1 31.4 ...
##  $ PctWhite            : num  81.8 89.2 90.9 91.7 94.1 ...
##  $ PctBlack            : num  2.595 0.969 0.74 0.783 0.27 ...
##  $ PctAsian            : num  4.822 2.246 0.466 1.161 0.666 ...
##  $ PctOtherRace        : num  1.843 3.741 2.747 1.363 0.492 ...
##  $ PctMarriedHouseholds: num  52.9 45.4 54.4 51 54 ...
##  $ BirthRate           : num  6.12 4.33 3.73 4.6 6.8 ...
##  $ deathRate           : num  165 161 175 195 144 ...
```

Data Quality

The dataset was inspected using the following methods:

1. Calucating the min, max, mean, median, 1st and 3rd quartile and identifying missing values (ie NA) using summary function
2. Visually inspect univariate data with histograms and boxplots
3. Scatterplots were completed for all independent variables against the deathrate. Note some scatterplots are not shown.

Missing Data

PctSomeCol18_24: The variable for percentage of residents with some college education, aged 18-24, had 152 missing values (NA) - 4.989% of the variable's rows. We will conduct the analysis with the observations with the data.

Tim: After thinking about this some more, I would vote to keep in this data with the caveat that there is a significant amount of data missing. the smaller subset might show something.

PctEmployed16_Over The variable for percentage of residents over 16 years old and employed has 2285 missing values (NA). This represents a very large portion of the overall dataset - 74.992% of the variable's rows. Due to the large number of missing values, this variable was not considered a key variable of the dataset.

Data suspected of being erroroneous

Median Age: There are 30 observations (0.98457%)with Median Age > 300. A median age of a county over 300 years of age is impossible in real life and these values were set to NA. This represents a small portion of the dataset.

Household size below: There are 61 counties reporting average household size less than 1 representing 2.002% of the observations. While it is reasonable to define a household as 1 or more persons living in the same occupancy space, a conclusive definition of a household was not provided with the dataset. For this reason in conjunction with the small percentage of affected rows, we kept these observations in the dataset.

avgAnnCount: There are 206 observations (6.76%) with the same value of 1962.667684. It is possible these data points are erroneous. Additionally, there are 150 observations greater than 2000, which is dispropritionately large in comparison to the bulk of values within the variable. The largest of these values are as large as >38,000 with 504 values > 10,000. Without further information about this variable, we can not conclusively state that the values are erroneous. For this reason we kept these observations in the dataset.

```r
# Dataframe created to count frequency of each county name.
GeoFreq <- as.data.frame(table(df$Geography))
GeoFreq[GeoFreq$Freq != 1, ]  # All counties listed only once
```

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

Univariate Analysis of Key Variables

Univariate analysis is shown in completness because of the issues described. It is important to have a strong understanding of the data.

Outcome deathRate and Average Annual Rate of Cancer

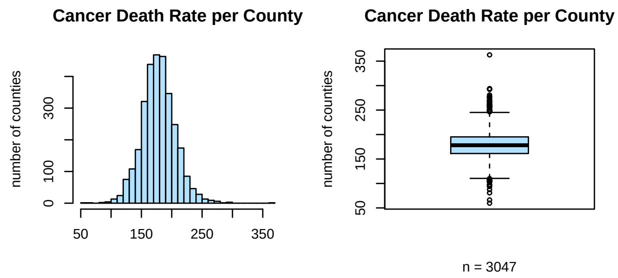
Summary Statistics:

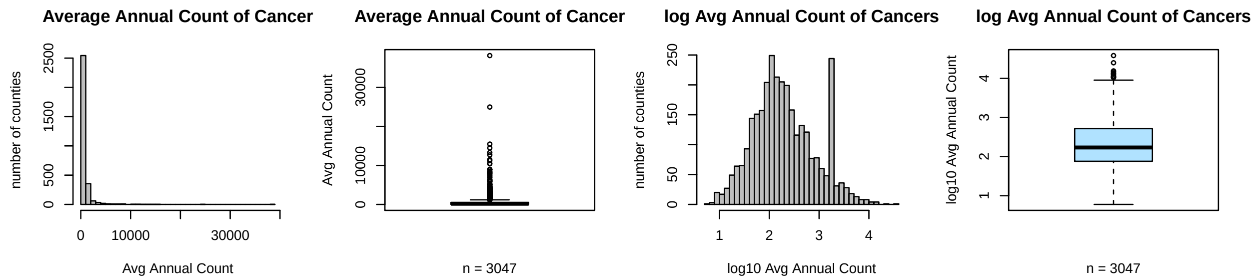
The cancer mortality rates are normally distributed. However, the average annual count of cancer is extremely right skewed due do magnitude changes. To help visualize the distribution, a log transformation was performed and produced a normal distribution curve in the histogram except for a bar count

```r
summary(data.frame(df$deathRate, df$avgAnnCount))
```

```
##   df.deathRate   df.avgAnnCount
##  Min.   : 59.7   Min.   :    6.0
##  1st Qu.:161.2   1st Qu.:   76.0
##  Median :178.1   Median :  171.0
##  Mean   :178.7   Mean   :  606.3
##  3rd Qu.:195.2   3rd Qu.:  518.0
##  Max.   :362.8   Max.   :38150.0
```

```r
layout(matrix(c(1, 2, 0, 0, 3, 4, 5, 6), 2, 4, byrow = T))  #Switched this to separate incidence and mo
color = c("lightskyblue1", "gray")

title = "Cancer Death Rate per County"
hist(df$deathRate, main = title, breaks = 30, ylab = "number of counties", xlab = "",
    col = color[1])
boxplot(df$deathRate, main = title, col = color, ylab = "number of counties", xlab = paste(c("n = ",
    sum(df$deathRate > 0, na.rm = TRUE)), collapse = ""))

title = "Average Annual Count of Cancer"
hist(df$avgAnnCount, main = title, breaks = 30, ylab = "number of counties", xlab = "Avg Annual Count",
    col = color[2])
boxplot(df$avgAnnCount, main = title, col = color, ylab = "Avg Annual Count", xlab = paste(c("n = ",
    sum(df$avgAnnCount > 0, na.rm = TRUE)), collapse = ""))

title = "log Avg Annual Count of Cancers"
hist(log(df$avgAnnCount, 10), main = title, breaks = 30, ylab = "number of counties",
    xlab = "log10 Avg Annual Count", col = color[2])
boxplot(log(df$avgAnnCount, 10), main = title, col = color, ylab = "log10 Avg Annual Count",
    xlab = paste(c("n = ", sum(log(df$avgAnnCount, 10) > 0, na.rm = TRUE)), collapse = ""))
```





Population and birth

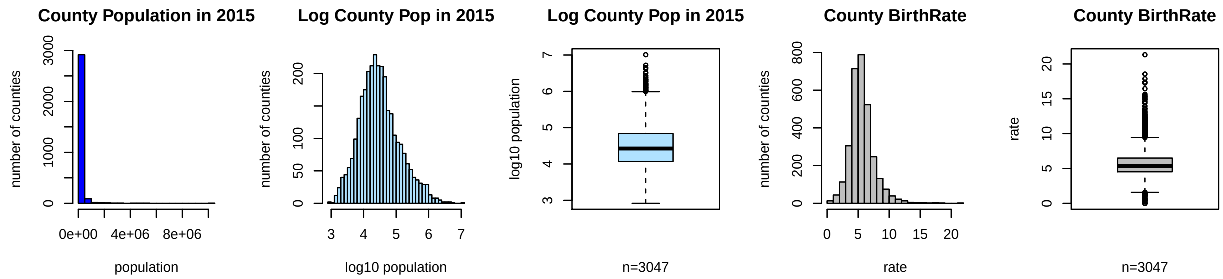
The counties ranged in population of 827 from Golden Valley County, Montana to 10170292 in Los Angeles County, California. There 42 counties that are have populations greater than 1 million and this produces an extreme right skew when the population is graphed on a histogram. Log transformations of the poulation helps visualize a distibution curve that looks normal. The birth rate didn't raise any concerns partly because we do not have the units to fully assess this variable.

Summary Statistics:

```r
summary(data.frame(df$popEst2015, df$BirthRate))
```

```
##  df.popEst2015       df.BirthRate
##  Min.   :     827   Min.   : 0.000
##  1st Qu.:   11684   1st Qu.: 4.521
##  Median :   26643   Median : 5.381
##  Mean   :  102637   Mean   : 5.640
##  3rd Qu.:   68671   3rd Qu.: 6.494
##  Max.   :10170292   Max.   :21.326
```

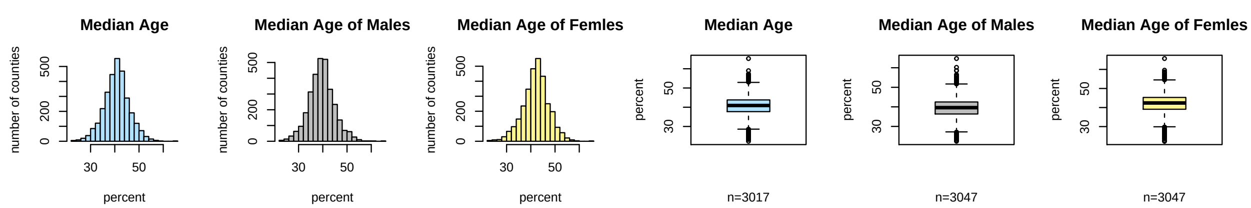
```r
par(mfrow = c(1, 5))
title = c("County Population in 2015", "Log County Pop in 2015", "County BirthRate")
y = c("number of counties", "log10 population", "population", "rate")
x = c("population", "log10 population", "rate")
hist(df$popEst2015, main = title[1], breaks = 30, ylab = y[1], xlab = x[1], col = "blue")
hist(log(df$popEst2015, 10), main = title[2], breaks = 30, ylab = y[1], xlab = x[2],
    col = "lightskyblue1")
boxplot(log(df$popEst2015, 10), main = title[2], col = "lightskyblue1", ylab = y[2],
    xlab = paste(c("n=", sum(df$popEst2015 >= 0, na.rm = TRUE)), collapse = ""))

hist(df$BirthRate, main = title[3], breaks = 30, ylab = y[1], xlab = x[3], col = "gray")
boxplot(df$BirthRate, main = title[3], col = "gray", ylab = y[4], xlab = paste(c("n=",
    sum(df$BirthRate >= 0, na.rm = TRUE)), collapse = ""))
```



Age

```r
# sum((df$MedianAge>300), na.rm=TRUE) # identifies 30 observations median age
# greater than 300
df$MedianAge_gr100 = 0  #creates a new column to save the dubious data
df$MedianAge_gr100[df$MedianAge > 300] <- df$MedianAge[df$MedianAge > 300]  # saves the dubious data
df$MedianAge[df$MedianAge > 300] <- NA  #update the Median Age
```

```r
par(mfrow = c(1, 6))
color = c("lightskyblue1", "gray", "khaki1", "darkseagreen1")
title = c("Median Age", "Median Age of Males", "Median Age of Femles")
y = c("number of counties", "percent")
x = c("percent")
hist(df$MedianAge, main = title[1], breaks = 30, ylab = y[1], xlab = x[1], col = color[1])
hist(df$MedianAgeMale, main = title[2], breaks = 30, ylab = y[1], xlab = x[1], col = color[2])
hist(df$MedianAgeFemale, main = title[3], breaks = 30, ylab = y[1], xlab = x[1],
    col = color[3])

boxplot(df$MedianAge, main = title[1], col = color[1], ylab = y[2], xlab = paste(c("n=",
    sum(df$MedianAge >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$MedianAgeMale, main = title[2], col = color[2], ylab = y[2], xlab = paste(c("n=",
    sum(df$MedianAgeMale >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$MedianAgeFemale, main = title[3], col = color[3], ylab = y[2], xlab = paste(c("n=",
    sum(df$MedianAgeFemale >= 0, na.rm = TRUE)), collapse = ""))
```



Ethnicity

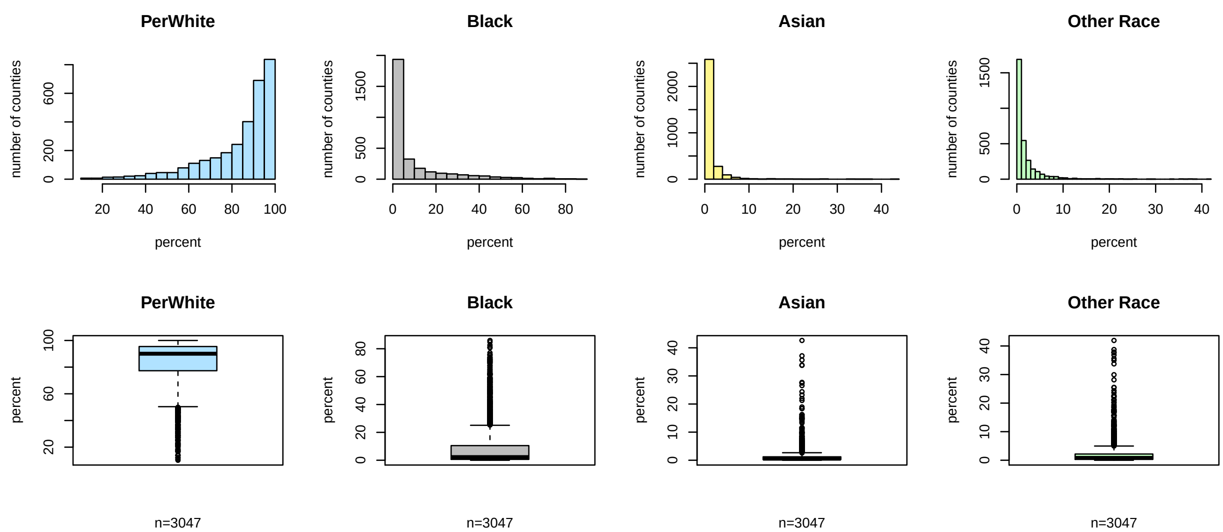
Summary Statistics:

```r
summary(data.frame(df$PctWhite, df$PctBlack, df$PctAsian, df$PctOtherRace), digits = 3)
```

```
##   df.PctWhite      df.PctBlack       df.PctAsian       df.PctOtherRace
##  Min.   : 10.2   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 77.3   1st Qu.: 0.621   1st Qu.: 0.254   1st Qu.: 0.295
```

```
##  Median : 90.1   Median : 2.248   Median : 0.550   Median : 0.826
##  Mean   : 83.6   Mean   : 9.108   Mean   : 1.254   Mean   : 1.984
##  3rd Qu.: 95.5   3rd Qu.:10.510   3rd Qu.: 1.221   3rd Qu.: 2.178
##  Max.   :100.0   Max.   :85.948   Max.   :42.619   Max.   :41.930
```

```
par(mfrow = c(2, 4))
color = c("lightskyblue1", "gray", "khaki1", "darkseagreen1")
title = c("PerWhite", "Black", "Asian", "Other Race")
y = c("number of counties", "percent")
x = c("percent")
hist(df$PctWhite, main = title[1], breaks = 30, ylab = y[1], xlab = x[1], col = color[1])
hist(df$PctBlack, main = title[2], breaks = 30, ylab = y[1], xlab = x[1], col = color[2])
hist(df$PctAsian, main = title[3], breaks = 30, ylab = y[1], xlab = x[1], col = color[3])
hist(df$PctOtherRace, main = title[4], breaks = 30, ylab = y[1], xlab = x[1], col = color[4])

boxplot(df$PctWhite, main = title[1], col = color[1], ylab = y[2], xlab = paste(c("n=",
    sum(df$PctWhite >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctBlack, main = title[2], col = color[2], ylab = y[2], xlab = paste(c("n=",
    sum(df$PctBlack >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctAsian, main = title[3], col = color[3], ylab = y[2], xlab = paste(c("n=",
    sum(df$PctAsian >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctOtherRace, main = title[4], col = color[4], ylab = y[2], xlab = paste(c("n=",
    sum(df$PctOtherRace >= 0, na.rm = TRUE)), collapse = ""))
```



Household and Marital status

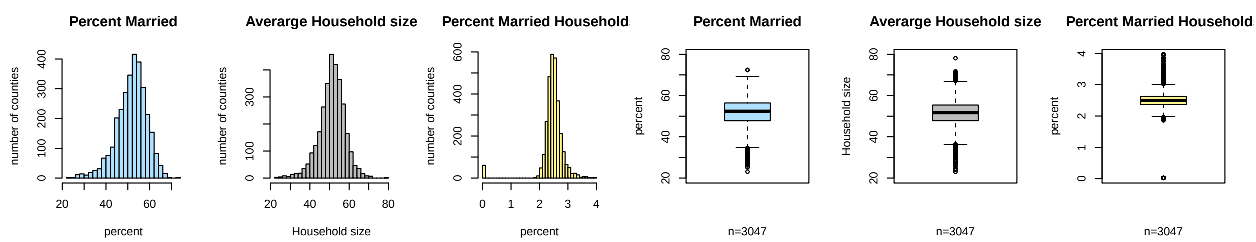
There are 61 counties reporting average household size less than 1 representing 2.002% of the observations. While it is reasonable to define a household as 1 or more persons living in the same occupancy space, a conclusive definition of a household was not provided with the dataset. For this reason in conjunction with the small percentage of affected rows, we left these observations in the dataset.

Summary Statistics:

```
summary(data.frame(df$PercentMarried, df$AvgHouseholdSize, df$PctMarriedHouseholds),
    digits = 3)
```

```
##  df.PercentMarried df.AvgHouseholdSize df.PctMarriedHouseholds
##  Min.   :23.1      Min.   :0.0221      Min.   :23.0
##  1st Qu.:47.8      1st Qu.:2.3700      1st Qu.:47.8
##  Median :52.4      Median :2.5000      Median :51.7
##  Mean   :51.8      Mean   :2.4797      Mean   :51.2
##  3rd Qu.:56.4      3rd Qu.:2.6300      3rd Qu.:55.4
##  Max.   :72.5      Max.   :3.9700      Max.   :78.1
```

```r
par(mfrow = c(1, 6))
color = c("lightskyblue1", "gray", "khaki1", "darkseagreen1")
title = c("Percent Married", "Averarge Household size", "Percent Married Households")
y = c("number of counties", "percent", "Household size")
x = c("percent", "Household size")
hist(df$PercentMarried, main = title[1], breaks = 30, ylab = y[1], xlab = x[1], col = color[1])
hist(df$PctMarriedHouseholds, main = title[2], breaks = 30, ylab = y[1], xlab = x[2],
    col = color[2])
hist(df$AvgHouseholdSize, main = title[3], breaks = 30, ylab = y[1], xlab = x[1],
    col = color[3])
yl = c(20, 80)
boxplot(df$PercentMarried, main = title[1], col = color[1], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PercentMarried >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctMarriedHouseholds, main = title[2], col = color[2], ylab = y[3], ylim = yl,
    xlab = paste(c("n=", sum(df$PctMarriedHouseholds >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$AvgHouseholdSize, main = title[3], col = color[3], ylab = y[2], xlab = paste(c("n=",
    sum(df$AvgHouseholdSize >= 0, na.rm = TRUE)), collapse = ""))
```



Education

The variables related to education can roughly be grouped as:

1. Some high school education (PctNoHS18_24)
2. High school completed (PctHS18_24, PctHS25_Over)
3. Some college education (PctSomeCol18_24)
4. Bachelor's degree completed (PctBachDeg18_24, PctBachDeg25_Over)

As stated previously, the only variable for *some college education*, PctSomeCol18_24, had a significant portion of data missing, and thus will not be considered in this analysis.

Summary Statistics:

```r
summary(data.frame(df$PctNoHS18_24, df$PctHS18_24, df$PctHS25_Over, df$PctBachDeg18_24,
    df$PctBachDeg25_Over), digits = 4)
```

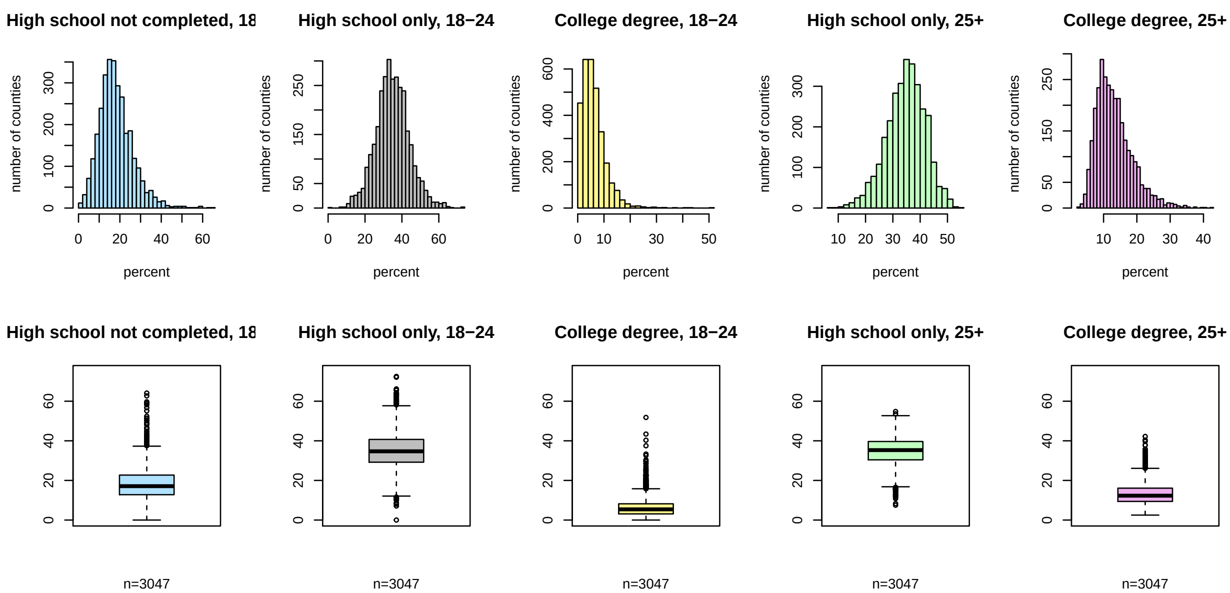
```
##  df.PctNoHS18_24 df.PctHS18_24  df.PctHS25_Over df.PctBachDeg18_24
##  Min.   : 0.00   Min.   : 0.0   Min.   : 7.50   Min.   : 0.000
##  1st Qu.:12.80   1st Qu.:29.2   1st Qu.:30.40   1st Qu.: 3.100
##  Median :17.10   Median :34.7   Median :35.30   Median : 5.400
```

```
##  Mean   :18.22   Mean   :35.0   Mean   :34.80   Mean   : 6.158
##  3rd Qu.:22.70   3rd Qu.:40.7   3rd Qu.:39.65   3rd Qu.: 8.200
##  Max.   :64.10   Max.   :72.5   Max.   :54.80   Max.   :51.800
##  df.PctBachDeg25_Over
##  Min.   : 2.50
##  1st Qu.: 9.40
##  Median :12.30
##  Mean   :13.28
##  3rd Qu.:16.10
##  Max.   :42.20
```

The basic summary statistics seem plausible and don't contain any concerning results. As expected, the percentage of county residents without a highschool degree is relatively low with a median of 17%, and roughly half of the percentage as those that have completed high school. It was also interesting to note that the interquartile range (IQR) was very similar for *PctHS18_24* and *PctHS25_Over* with the primary difference being at the extreme ends - the 18-24 range went as low as 0% and a max of 72.%, while the over 25 age group more centered with a minimum of 7.5% and a max of 54%. This should be expected because the over-25 age group likely includes more people, as well as the extra time and opportunity to complete a high school degree.

The following histogram plots show the distributions of the education variables.

```r
par(mfrow = c(2, 5))
color = c("lightskyblue1", "gray", "khaki1", "darkseagreen1", "plum2")
title = c("High school not completed, 18-24", "High school only, 18-24", "College degree, 18-24",
    "High school only, 25+", "College degree, 25+")
y = c("number of counties")
x = c("percent")
yl = c(0, 75)
hist(df$PctNoHS18_24, main = title[1], breaks = 30, ylab = y[1], xlab = x[1], col = color[1])
hist(df$PctHS18_24, main = title[2], breaks = 30, ylab = y[1], xlab = x[1], col = color[2])
hist(df$PctBachDeg18_24, main = title[3], breaks = 30, ylab = y[1], xlab = x[1],
    col = color[3])
hist(df$PctHS25_Over, main = title[4], breaks = 30, ylab = y[1], xlab = x[1], col = color[4])
hist(df$PctBachDeg25_Over, main = title[5], breaks = 30, ylab = y[1], xlab = x[1],
    col = color[5])

boxplot(df$PctNoHS18_24, main = title[1], col = color[1], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctNoHS18_24 >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctHS18_24, main = title[2], col = color[2], ylab = y[2], ylim = yl, xlab = paste(c("n=",
    sum(df$PctHS18_24 >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctBachDeg18_24, main = title[3], col = color[3], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctBachDeg18_24 >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctHS25_Over, main = title[4], col = color[4], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctBachDeg18_24 >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctBachDeg25_Over, main = title[5], col = color[5], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctBachDeg25_Over >= 0, na.rm = TRUE)), collapse = ""))
```



Of the five variables considered, only *PctHS18_24* showed a normal distribution. The other variables had some skew, either positive or negative. The variable *PctNoHS18_24*, no high school education, for example showed a left skew meaning that smaller percentages, 10-20%, were much more common than high percentages, 30+%. Conversely, the distribution of county residents that only had a high school education, *PctHS25_Over*, had a postive right skew, with highly percentages much more common than low percentages.

The variable *PctBachDeg18_24* was the most extreme distribution with a far-left skew. This is likely due to the deinition of the variable. Most people start college at age 18 and finish by age 22-23. Because the age window was 18-24, the majority of considered residents would not have had an opportunity to yet finish college. This observed skew is less pronounced for the variable *PctBachDeg25_Over*, which includes all county residents over the age of 25. For this reason, *PctBachDeg18_24* will not be considered a key variable.

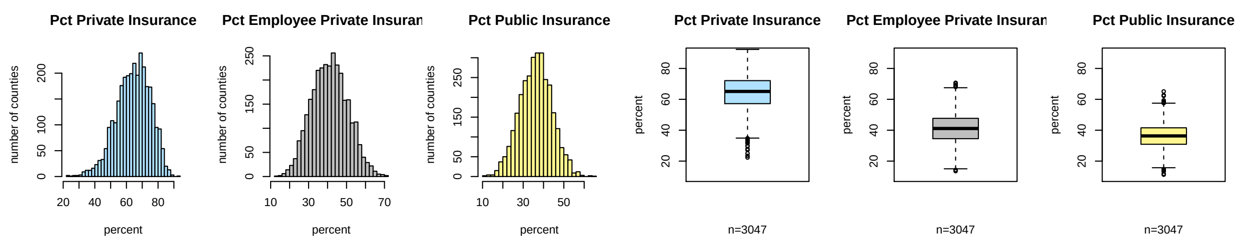
Insurance Coverage

Summary Statistics:

```
summary(data.frame(df$PctNoHS18_24, df$PctHS18_24, df$PctHS25_Over, df$PctBachDeg18_24,
    df$PctBachDeg25_Over), digits = 4)
```

```
##  df.PctNoHS18_24 df.PctHS18_24  df.PctHS25_Over df.PctBachDeg18_24
##  Min.   : 0.00   Min.   : 0.0   Min.   : 7.50   Min.   : 0.000
##  1st Qu.:12.80   1st Qu.:29.2   1st Qu.:30.40   1st Qu.: 3.100
##  Median :17.10   Median :34.7   Median :35.30   Median : 5.400
##  Mean   :18.22   Mean   :35.0   Mean   :34.80   Mean   : 6.158
##  3rd Qu.:22.70   3rd Qu.:40.7   3rd Qu.:39.65   3rd Qu.: 8.200
##  Max.   :64.10   Max.   :72.5   Max.   :54.80   Max.   :51.800
##  df.PctBachDeg25_Over
##  Min.   : 2.50
##  1st Qu.: 9.40
##  Median :12.30
##  Mean   :13.28
##  3rd Qu.:16.10
##  Max.   :42.20
```

```r
par(mfrow = c(1, 6))
color = c("lightskyblue1", "gray", "khaki1", "darkseagreen1")
title = c("Pct Private Insurance", "Pct Employee Private Insurance", "Pct Public Insurance")
y = c("number of counties", "percent")
x = c("percent")
yl = c(10, 90)
hist(df$PctPrivateCoverage, main = title[1], breaks = 30, ylab = y[1], xlab = x[1],
    col = color[1])
hist(df$PctEmpPrivCoverage, main = title[2], breaks = 30, ylab = y[1], xlab = x[1],
    col = color[2])
hist(df$PctPublicCoverage, main = title[3], breaks = 30, ylab = y[1], xlab = x[1],
    col = color[3])

boxplot(df$PctPrivateCoverage, main = title[1], col = color[1], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctPrivateCoverage >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctEmpPrivCoverage, main = title[2], col = color[2], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctEmpPrivCoverage >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctPublicCoverage, main = title[3], col = color[3], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctPublicCoverage >= 0, na.rm = TRUE)), collapse = ""))
```



Employment

Other than the 152 missing data observations, the two variables describing employment of residents 16 years or older show a normal distribution.

Summary Statistics:

```r
summary(data.frame(df$PctEmployed16_Over, df$PctUnemployed16_Over), digits = 3)
```

```
##   df.PctEmployed16_Over df.PctUnemployed16_Over
##   Min.   :17.6          Min.   : 0.40
##   1st Qu.:48.6          1st Qu.: 5.50
##   Median :54.5          Median : 7.60
##   Mean   :54.2          Mean   : 7.85
##   3rd Qu.:60.3          3rd Qu.: 9.70
##   Max.   :80.1          Max.   :29.40
##   NA's   :152
```

```r
par(mfrow = c(1, 4))
color = c("lightskyblue1", "gray", "khaki1", "darkseagreen1", "plum2")
title = c("Employed 16 yrs or older", "Unemployed 16 yrs or older")
y = c("number of counties", "percent")
x = c("percent")
yl = c(0, 80)
hist(df$PctEmployed16_Over, main = title[1], breaks = 30, ylab = y[1], xlab = x[1],
```

```r
    col = color[1])
hist(df$PctHS18_24, main = title[2], breaks = 30, ylab = y[1], xlab = x[1], col = color[2])

boxplot(df$PctEmployed16_Over, main = title[1], col = color[1], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctEmployed16_Over >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$PctHS18_24, main = title[2], col = color[2], ylab = y[2], ylim = yl, xlab = paste(c("n=",
    sum(df$PctHS18_24 >= 0, na.rm = TRUE)), collapse = ""))
```

Income and Poverty level

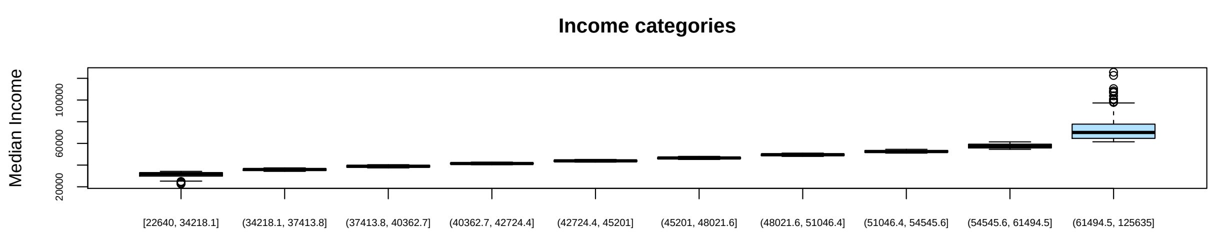
Median income was provided for each county and assigned a bin containing 302-306 observations. We confirmed that the 10 income bins are based on the deciles of the median income of the dataset. Median income has a right skew. Percent of residents below the poverty level was also included the dataset and presents a right skew as well.

The binned income

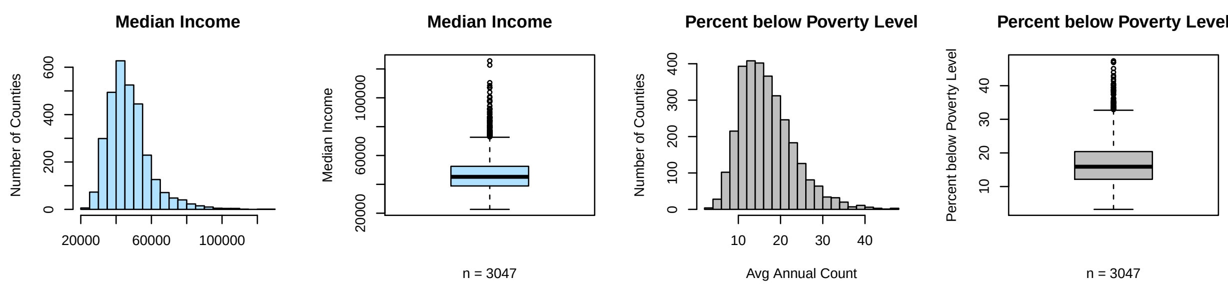
```r
df$binnedInc <- factor(df$binnedInc, levels = c("[22640, 34218.1]", "(34218.1, 37413.8]",
    "(37413.8, 40362.7]", "(40362.7, 42724.4]", "(42724.4, 45201]", "(45201, 48021.6]",
    "(48021.6, 51046.4]", "(51046.4, 54545.6]", "(54545.6, 61494.5]", "(61494.5, 125635]"))
summary(df$binnedInc)
```

```
##   [22640, 34218.1] (34218.1, 37413.8] (37413.8, 40362.7]
##                306                304                304
## (40362.7, 42724.4]   (42724.4, 45201]   (45201, 48021.6]
##                304                305                306
## (48021.6, 51046.4] (51046.4, 54545.6] (54545.6, 61494.5]
##                305                305                306
##  (61494.5, 125635]
##                302
```

```r
boxplot(df$medIncome ~ df$binnedInc, data = df, cex.axis = 0.58, main = "Income categories",
    ylab = "Median Income", col = "lightskyblue1")
```



```r
par(mfrow = c(1, 4))
color = "lightskyblue1"

title = "Median Income"
hist(df$medIncome, main = title, breaks = 30, ylab = "Number of Counties", xlab = "",
    col = color)
boxplot(df$medIncome, main = title, col = color, ylab = "Median Income", xlab = paste(c("n = ",
    sum(df$medIncome >= 0, na.rm = TRUE)), collapse = ""))

color = "gray"
title = "Percent below Poverty Level"
hist(df$povertyPercent, main = title, breaks = 30, ylab = "Number of Counties", xlab = "Avg Annual Coun
    col = color)
boxplot(df$povertyPercent, main = title, col = color, ylab = "Percent below Poverty Level",
    xlab = paste(c("n = ", sum(df$povertyPercent > 0, na.rm = TRUE)), collapse = ""))
```



Analysis of Key Relationships

Table below show below shows the correlation coeffiencients for r

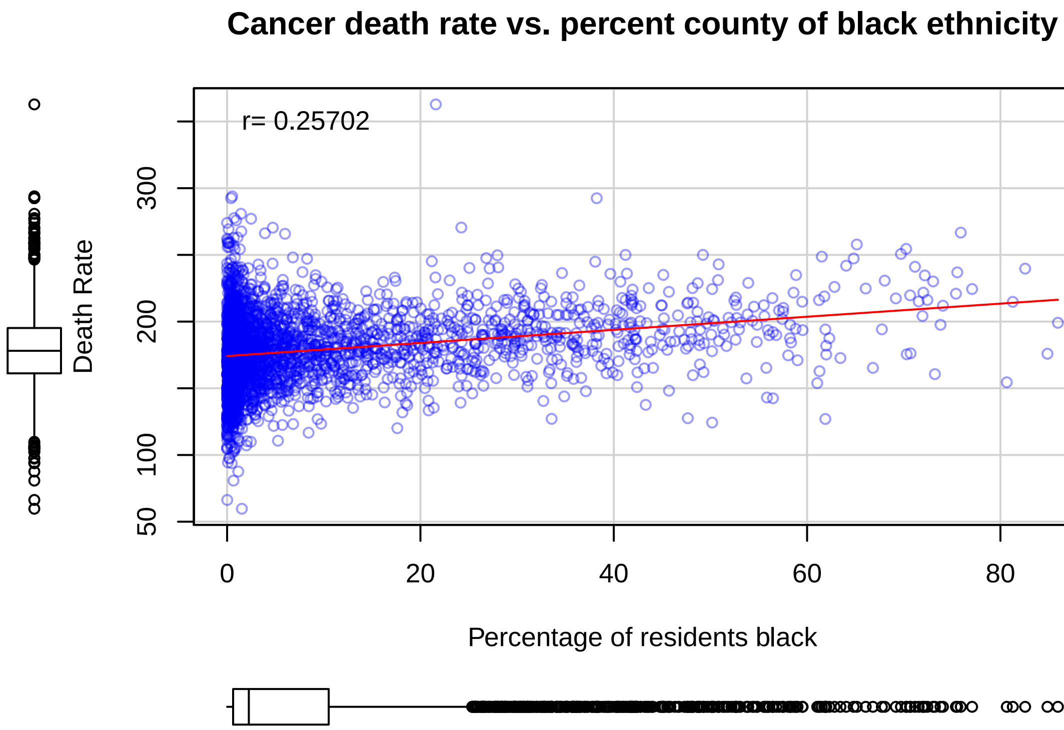
```r
x <- names(df)[!names(df) %in% c("deathRate", "X")]

vars <- c()
cors <- c()

for (i in x) {
    if (class(df[, i]) == "numeric") {
        vars <- c(vars, i)
        cors <- c(cors, round(cor(df$deathRate, df[, i], use = "complete.obs"), digits = 4))
    }
}

cor_df <- data.frame(vars, cors, row.names = NULL)
cor_df <- cor_df[order(abs(cor_df$cors), decreasing = T), ]

kable(cor_df, col.names = c("Variable", "Correlation Coefficient"), row.names = F,
    caption = "Correlation Coefficients of Indepent Variables to Death Rate")
```

Table 2: Correlation Coefficients of Indepent Variables to Death Rate

Variable	Correlation Coefficient
PctBachDeg25_Over	-0.4855
povertyPercent	0.4294
PctEmployed16_Over	-0.4120
PctHS25_Over	0.4046
PctPublicCoverage	0.4046
PctPrivateCoverage	-0.3861
PctUnemployed16_Over	0.3784
PctMarriedHouseholds	-0.2933
PctBachDeg18_24	-0.2878
PctEmpPrivCoverage	-0.2674
PercentMarried	-0.2668
PctHS18_24	0.2620
PctBlack	0.2570
PctOtherRace	-0.1899
PctSomeCol18_24	-0.1887
PctAsian	-0.1863
PctWhite	-0.1774
avgAnnCount	-0.1435
PctNoHS18_24	0.0885
BirthRate	-0.0874
AvgHouseholdSize	-0.0369
MedianAgeMale	-0.0219
MedianAgeFemale	0.0120
MedianAge_gr100	0.0050
MedianAge	-0.0043

MIKE TO ADD HIS SECTION HERE

```
r_cor = round(cor(df$deathRate, df$PctBlack, use = "complete.obs"), 5)
scatterplot(df$PctBlack, df$deathRate, xlab = "Percentage of residents black", ylab = "Death Rate",
    main = "Cancer death rate vs. percent county of black ethnicity", legend("topleft",
        bty = "n", legend = paste("r=", r_cor)), col = c("red", "green", rgb(0, 0,
        250, 100, maxColorValue = 255)))
```



END MIKE SECITON

Household size and Marital status

Average household size did not have any correlation with death rate. However, percent married and percent married households did.

```
# sub <- cor_df[cor_df$vars %in% c('povertyPercent'),] kable(sub, row.names=F,
# caption='Correlation Coefficients of Indepent Variables to Death Rate',
# col.names=c('Variable', 'Correlation Coefficient'))
```

Independent variable	cor to death rate
PercentMarried	-0.2668205
PctMarriedHouseholds	-0.2933253
AvgHouseholdSize	-0.0369053

Education

To begin, a correlation table of the education variables is shown vs. the target variable *deathRate*

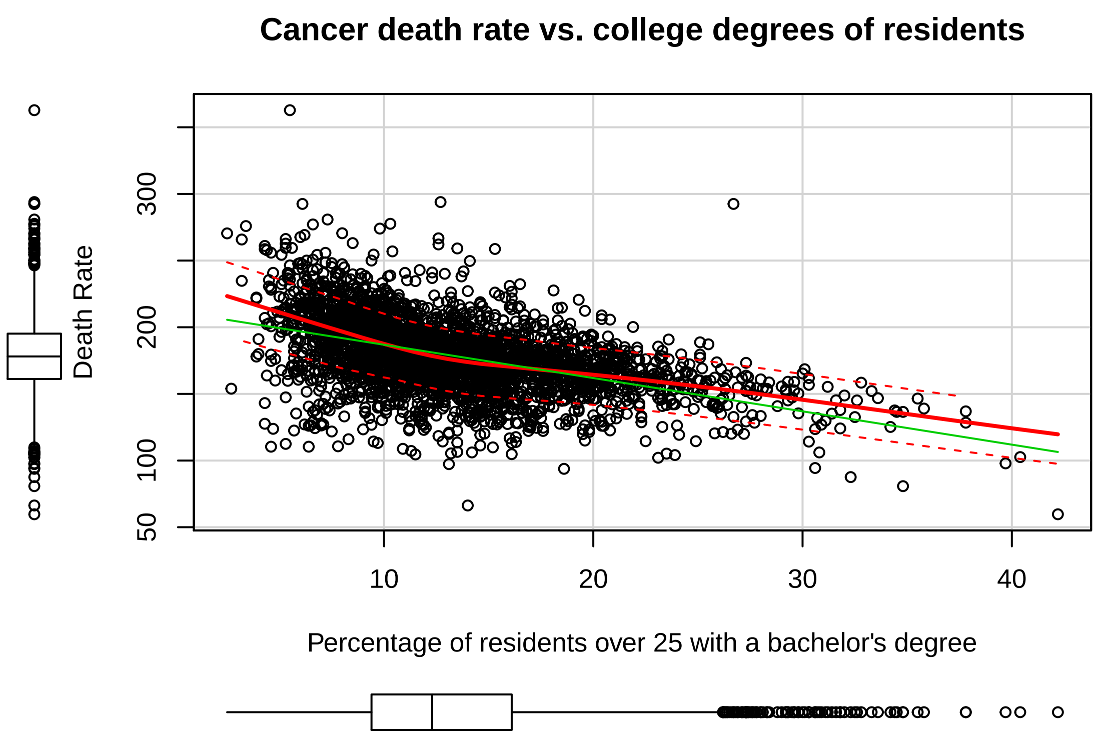
Independent variable	cor to death rate
PctNoHS18_24	0.0884626
PctHS18_24	0.2619759
PctSomeCol18_24	-0.1886877
PctBachDeg18_24	-0.2878174

Independent variable	cor to death rate
PctHS25_Over	0.4045891
PctBachDeg25_Over	-0.4854773

The first thing to note is that the correlations to death rate of the 18-24 age variables were generally weaker compared to the broader “over 25” age variables. One weakness of the dataset is that the absolute quantities of these variables are unknown - only the percentage is known. For that reason, it may be safe to presume that the 18-24 age variables are prone to more variation and worse correlation to the dependent variable, death rate.

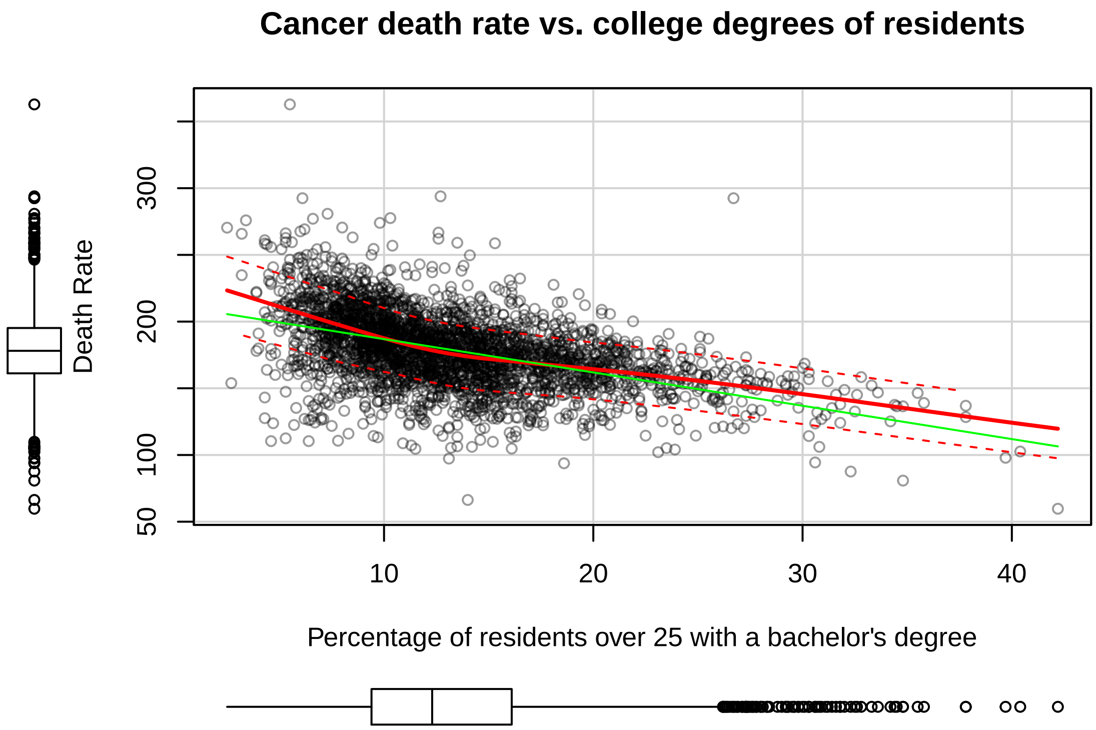
Observing only the two “over 25 age” variables, we see two distinctly different trends. The first - percentage of residents with a college degree - shows a moderately decreasing trend with a negative correation of -0.4854773.

```
scatterplot(df$PctBachDeg25_Over, df$deathRate, xlab = "Percentage of residents over 25 with a bachelor
    ylab = "Death Rate", main = "Cancer death rate vs. college degrees of residents")
```



The second variable, *PctHS25_Over*, conversely shows an increasing trend with a correlation of 0.4045891.

```
r_cor = round(cor(df$deathRate, df$PctHS25_Over, use = "complete.obs"), 5)
scatterplot(df$PctBachDeg25_Over, df$deathRate, xlab = "Percentage of residents over 25 with a bachelor
    col = c("green1", "red", rgb(0, 0, 0, 100, maxColorValue = 255)), ylab = "Death Rate",
    main = "Cancer death rate vs. college degrees of residents")
```



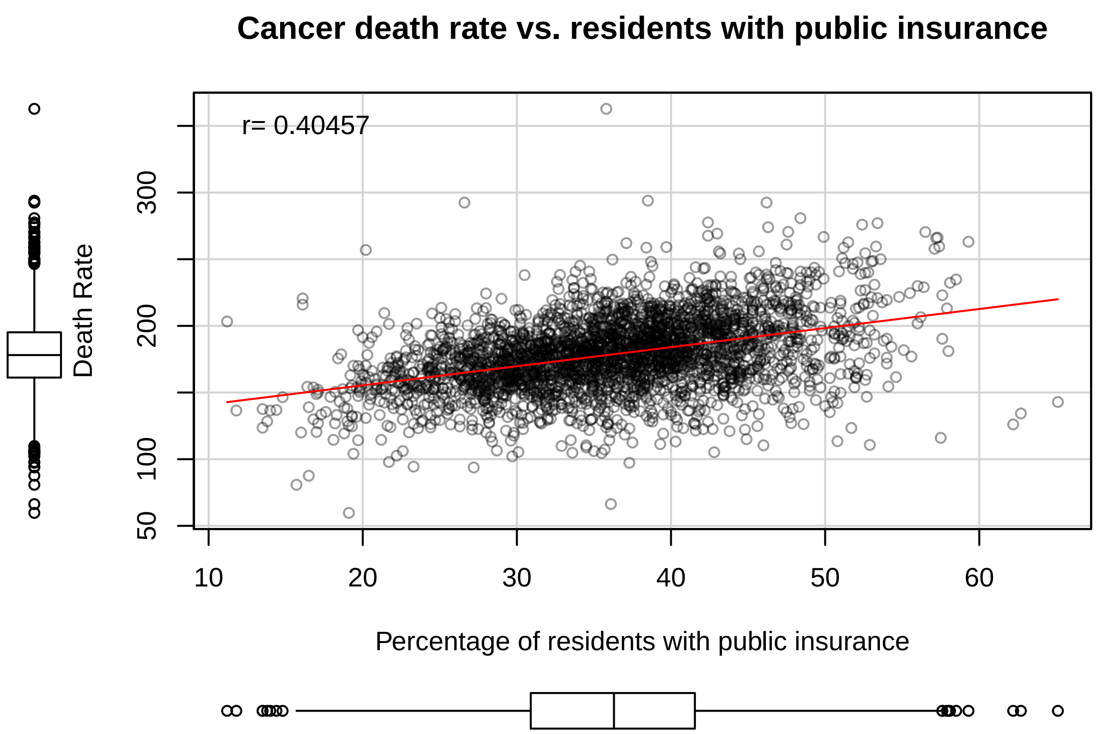
Overall, these two variables agree with each other: when the percentage of county residents have more education, the mortality rate due to cancer is generally lower.

Insurance

Death rate Scatterplot and Correlations by

Independent variable	cor to death rate
PctPrivateCoverage	-0.3860655
PctEmpPrivCoverage	-0.2673994
PctPublicCoverage	0.4045717

```
r_cor = round(cor(df$deathRate, df$PctPublicCoverage, use = "complete.obs"), 5)
scatterplot(df$PctPublicCoverage, df$deathRate, xlab = "Percentage of residents with public insurance",
    ylab = "Death Rate", main = "Cancer death rate vs. residents with public insurance",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))
```

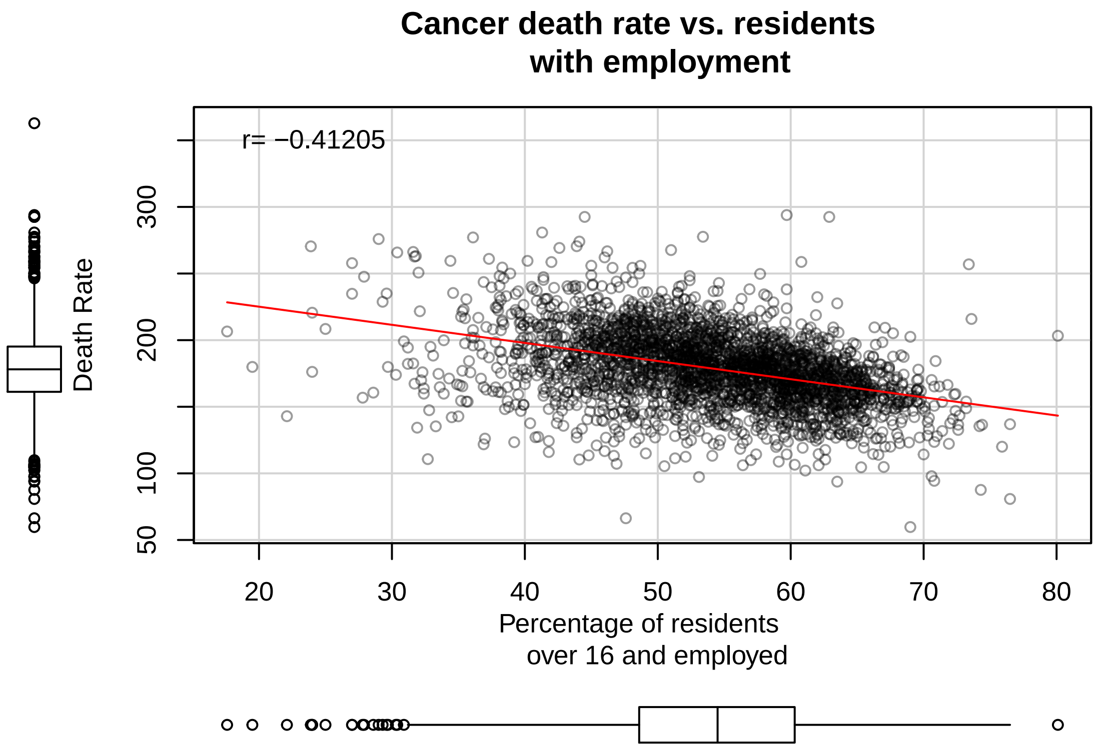


Employment

16 years or older

Independent variable	cor to death rate
PctEmployed16_Over	-0.4120458
PctUnemployed16_Over	0.3784124

```
r_cor = round(cor(df$deathRate, df$PctEmployed16_Over, use = "complete.obs"), 5)
scatterplot(df$PctEmployed16_Over, df$deathRate, xlab = "Percentage of residents
    over 16 and employed",
    ylab = "Death Rate", main = "Cancer death rate vs. residents
    with employment",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))
```

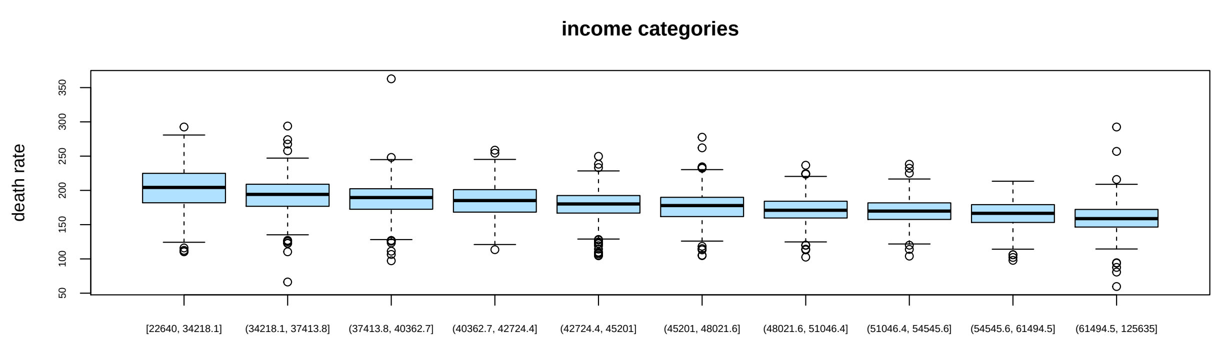


Income

The binned income shows a downward trend in death rate as the binned median incomes increases.

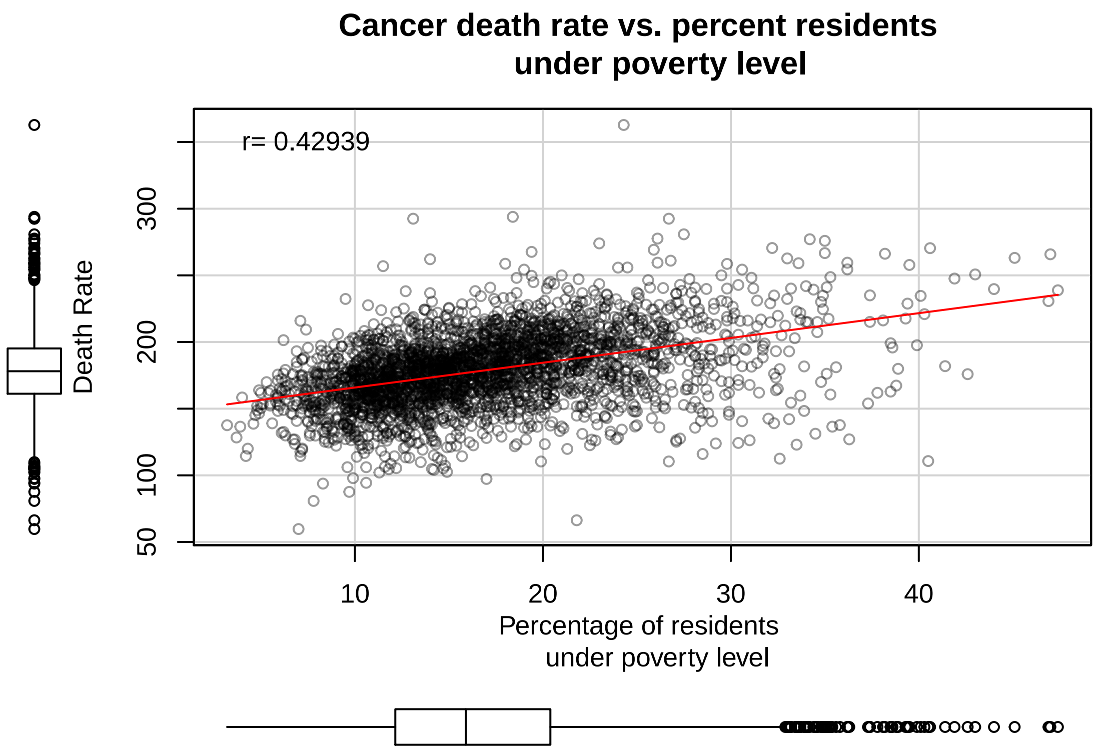
Median income

```r
boxplot(df$deathRate ~ df$binnedInc, data = df, cex.axis = 0.58, main = "income categories",
    ylab = "death rate", col = "lightskyblue1")
```

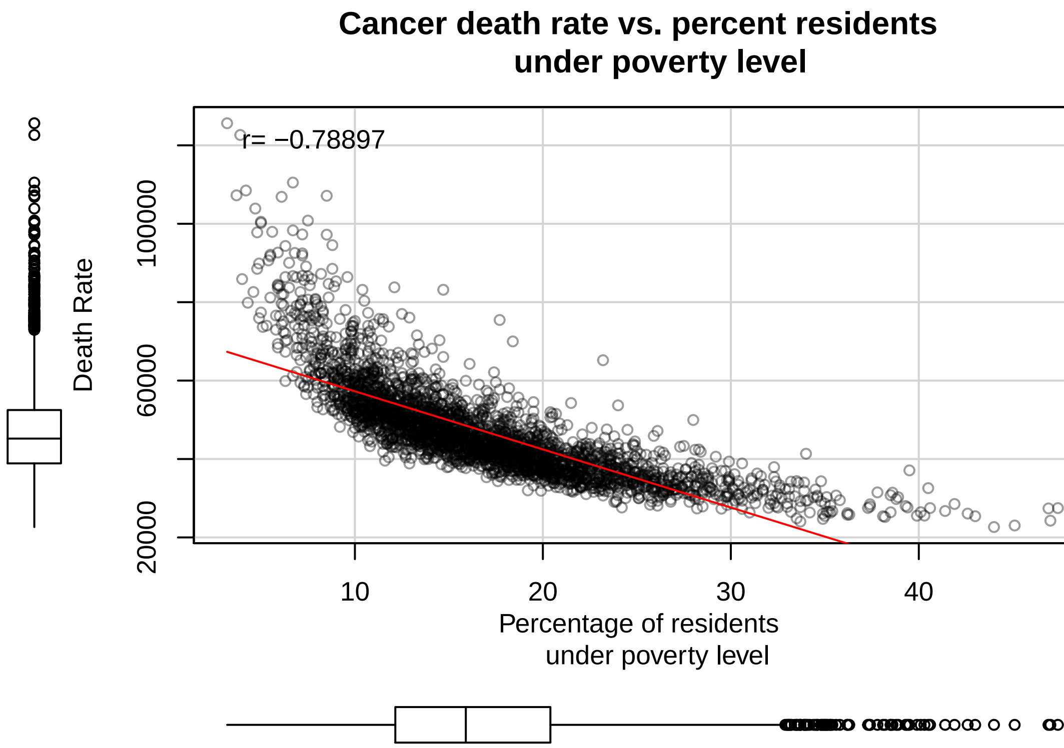


Independent variable	cor to death rate
povertyPercent	0.429389
medIncome	-0.4286149

```r
r_cor = round(cor(df$deathRate, df$povertyPercent, use = "complete.obs"), 5)
scatterplot(df$povertyPercent, df$deathRate, xlab = "Percentage of residents
    under poverty level",
    ylab = "Death Rate", main = "Cancer death rate vs. percent residents
    under poverty level",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))
```




```r
r_cor = round(cor(df$medIncome, df$povertyPercent, use = "complete.obs"), 5)
scatterplot(df$povertyPercent, df$medIncome, xlab = "Percentage of residents
    under poverty level",
    ylab = "Death Rate", main = "Cancer death rate vs. percent residents
    under poverty level",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))
```



Analysis Variables with high correclation and Secondary Effects

```r
# Tim place the heat map correlation coefficients map here



cor(df[, c("MedianAge", "PctBlack", "PctHS25_Over", "PctBachDeg25_Over", "PctPrivateCoverage",
    "PctEmpPrivCoverage", "PctPublicCoverage", "PctEmployed16_Over", "PctUnemployed16_Over")],
    use = "complete.obs")
```

```
##                        MedianAge   PctBlack PctHS25_Over PctBachDeg25_Over
## MedianAge             1.00000000 -0.2117756   0.32625378        -0.1467633
## PctBlack             -0.21177558  1.0000000  -0.02437170        -0.1447773
## PctHS25_Over          0.32625378 -0.0243717   1.00000000        -0.7423720
## PctBachDeg25_Over    -0.14676329 -0.1447773  -0.74237196         1.0000000
## PctPrivateCoverage    0.07223544 -0.3440071  -0.22558984         0.6065063
## PctEmpPrivCoverage   -0.23304808 -0.2334563  -0.22711435         0.5444610
## PctPublicCoverage     0.42175048  0.1912806   0.42436319        -0.6353066
## PctEmployed16_Over   -0.19251036 -0.3379401  -0.34656204         0.6165347
## PctUnemployed16_Over -0.12659162  0.4733857   0.08380191        -0.3699384
##                      PctPrivateCoverage PctEmpPrivCoverage
## MedianAge                    0.07223544         -0.2330481
## PctBlack                    -0.34400706         -0.2334563
## PctHS25_Over                -0.22558984         -0.2271143
## PctBachDeg25_Over            0.60650626          0.5444610
## PctPrivateCoverage           1.00000000          0.8258710
## PctEmpPrivCoverage           0.82587105          1.0000000
## PctPublicCoverage           -0.72252854         -0.7827213
## PctEmployed16_Over           0.69878309          0.7021542
```

```
## PctUnemployed16_Over        -0.62993190        -0.4694935
##                      PctPublicCoverage PctEmployed16_Over
## MedianAge                    0.4217505         -0.1925104
## PctBlack                     0.1912806         -0.3379401
## PctHS25_Over                 0.4243632         -0.3465620
## PctBachDeg25_Over           -0.6353066          0.6165347
## PctPrivateCoverage          -0.7225285          0.6987831
## PctEmpPrivCoverage          -0.7827213          0.7021542
## PctPublicCoverage            1.0000000         -0.7704844
## PctEmployed16_Over          -0.7704844          1.0000000
## PctUnemployed16_Over         0.5306896         -0.6475222
##                      PctUnemployed16_Over
## MedianAge                     -0.12659162
## PctBlack                       0.47338569
## PctHS25_Over                   0.08380191
## PctBachDeg25_Over             -0.36993836
## PctPrivateCoverage            -0.62993190
## PctEmpPrivCoverage            -0.46949353
## PctPublicCoverage              0.53068959
## PctEmployed16_Over            -0.64752218
## PctUnemployed16_Over           1.00000000
```

Conclusion