W203 Lab 1

Eduard Gelman, Jennifer Mahle, Atit Wongnophadol

September 24, 2018

Background

We are tasked to understand factors that predict cancer mortality rates, with the ultimate aim of identifying communities for social interventions, and of understanding which interventions are likely to have the most impact on decreasing the death rate due to cancer-related deaths.

Objective

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

Introduction

State the research question that motivates your analysis.

For purposes of this EDA exercise, we will assume that we serve a task force to the Federal Insurance Commissioner's Office, evaluating potential policy interventions. Therefore, we would like to explore the relationships between incomes, insurance type (private vs. public) and cancer mortality rates. Ideally, our analysis explores insurance rates, which are addressable by this Office, and takes into account other potential drivers of cancer mortality rates that will not be changed by this Office's policies.

Since this is a cross section of all counties, we will perform descriptive statistics only. This analysis will have two functions: 1. help inform policy recommendations for the Federal Insurance Commissioner's Office 2. discuss future analytical work, including statistical modeling

Load your data set into R.

```
# read data
dataset <- read.csv("cancer.csv", sep=",", header = TRUE)
```

Describe your data set. What types of variables does it contain? How many observations are there?

The data set is a comma separated value (*.csv*) file that contains 3,047 rows and 30 columns. Each row contains cross sectional, county-level demographic information, including 29 variables that describe the income, age, education level, employment, race, birth, death rates and so forth, for the county. The specific variables, their data types (i.e., int, num, and Factor) are listed as follows (note that the first vector "X" is simply for the index of each row).

```r
# list all the vectors in the data set, with data type, row count, and data samples
str(dataset)
```

```
## '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 ...
```

Evaluate the data quality. Are there any issues with the data? Explain how you handled these potential issues.

We check for data integrity issue by (i) inspecting summary statistics for all variables in the data set (ii) inspecting data plots for numeric or integer variables.

```r
# (i) Data integrity check by summary statistics for all variables
summary(dataset)
```

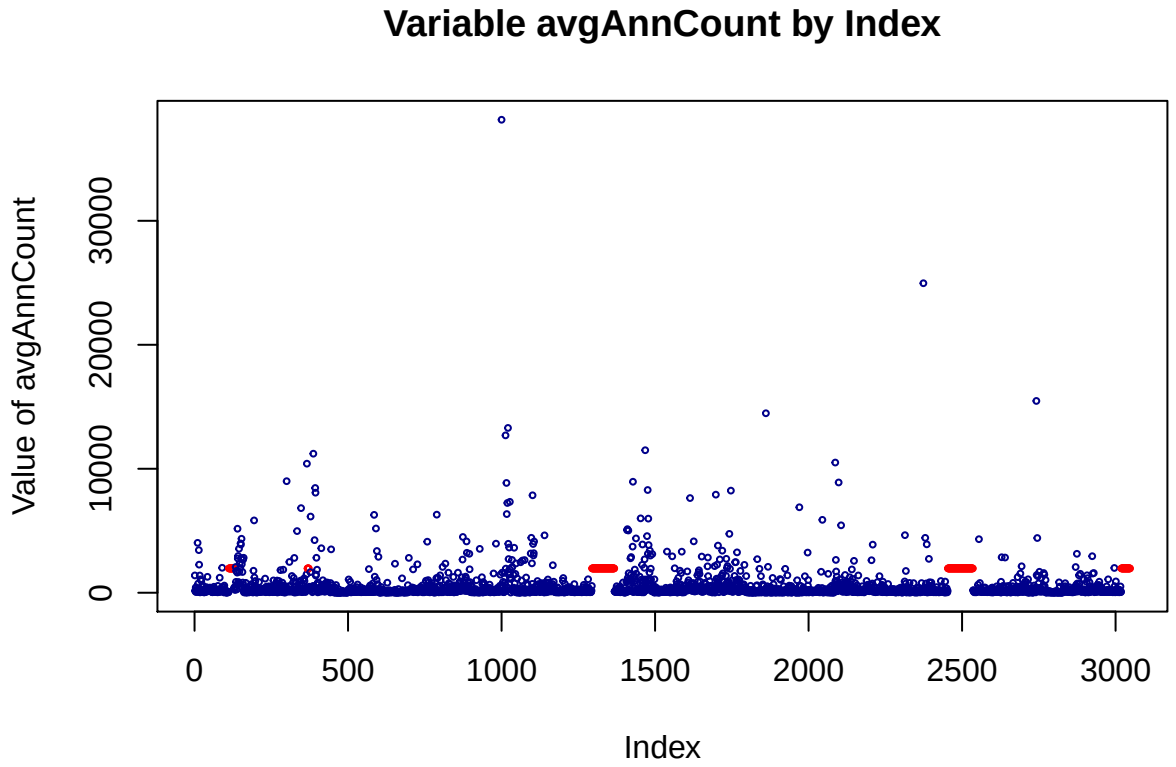
```
##        X           avgAnnCount        medIncome        popEst2015       povertyPercent
##  Min.   :   1.0   Min.   :    6.0   Min.   : 22640   Min.   :     827   Min.   : 3.20
##  1st Qu.: 762.5   1st Qu.:   76.0   1st Qu.: 38882   1st Qu.:   11684   1st Qu.:12.15
##  Median :1524.0   Median :  171.0   Median : 45207   Median :   26643   Median :15.90
##  Mean   :1524.0   Mean   :  606.3   Mean   : 47063   Mean   :  102637   Mean   :16.88
##  3rd Qu.:2285.5   3rd Qu.:  518.0   3rd Qu.: 52492   3rd Qu.:   68671   3rd Qu.:20.40
##  Max.   :3047.0   Max.   :38150.0   Max.   :125635   Max.   :10170292   Max.   :47.40
##
##                binnedInc      MedianAge      MedianAgeMale   MedianAgeFemale
##  (45201, 48021.6]  : 306   Min.   : 22.30   Min.   :22.40   Min.   :22.30
##  (54545.6, 61494.5]: 306   1st Qu.: 37.70   1st Qu.:36.35   1st Qu.:39.10
##  [22640, 34218.1]  : 306   Median : 41.00   Median :39.60   Median :42.40
##  (42724.4, 45201]  : 305   Mean   : 45.27   Mean   :39.57   Mean   :42.15
##  (48021.6, 51046.4]: 305   3rd Qu.: 44.00   3rd Qu.:42.50   3rd Qu.:45.30
##  (51046.4, 54545.6]: 305   Max.   :624.00   Max.   :64.70   Max.   :65.70
##  (Other)           :1214
##                            Geography    AvgHouseholdSize PercentMarried   PctNoHS18_24
##  Abbeville County, South Carolina:   1   Min.   :0.0221   Min.   :23.10   Min.   : 0.00
##  Acadia Parish, Louisiana        :   1   1st Qu.:2.3700   1st Qu.:47.75   1st Qu.:12.80
##  Accomack County, Virginia       :   1   Median :2.5000   Median :52.40   Median :17.10
##  Ada County, Idaho               :   1   Mean   :2.4797   Mean   :51.77   Mean   :18.22
##  Adair County, Iowa              :   1   3rd Qu.:2.6300   3rd Qu.:56.40   3rd Qu.:22.70
##  Adair County, Kentucky          :   1   Max.   :3.9700   Max.   :72.50   Max.   :64.10
##  (Other)                         :3041
##    PctHS18_24   PctSomeCol18_24 PctBachDeg18_24   PctHS25_Over   PctBachDeg25_Over
##  Min.   : 0.0   Min.   : 7.10   Min.   : 0.000   Min.   : 7.50   Min.   : 2.50
##  1st Qu.:29.2   1st Qu.:34.00   1st Qu.: 3.100   1st Qu.:30.40   1st Qu.: 9.40
##  Median :34.7   Median :40.40   Median : 5.400   Median :35.30   Median :12.30
##  Mean   :35.0   Mean   :40.98   Mean   : 6.158   Mean   :34.80   Mean   :13.28
##  3rd Qu.:40.7   3rd Qu.:46.40   3rd Qu.: 8.200   3rd Qu.:39.65   3rd Qu.:16.10
##  Max.   :72.5   Max.   :79.00   Max.   :51.800   Max.   :54.80   Max.   :42.20
##                 NA's   :2285
##  PctEmployed16_Over PctUnemployed16_Over PctPrivateCoverage PctEmpPrivCoverage PctPublicCoverage
##  Min.   :17.60      Min.   : 0.400       Min.   :22.30      Min.   :13.5       Min.   :11.20
##  1st Qu.:48.60      1st Qu.: 5.500       1st Qu.:57.20      1st Qu.:34.5       1st Qu.:30.90
##  Median :54.50      Median : 7.600       Median :65.10      Median :41.1       Median :36.30
##  Mean   :54.15      Mean   : 7.852       Mean   :64.35      Mean   :41.2       Mean   :36.25
##  3rd Qu.:60.30      3rd Qu.: 9.700       3rd Qu.:72.10      3rd Qu.:47.7       3rd Qu.:41.55
##  Max.   :80.10      Max.   :29.400       Max.   :92.30      Max.   :70.7       Max.   :65.10
##  NA's   :152
##     PctWhite         PctBlack          PctAsian        PctOtherRace     PctMarriedHouseholds
##  Min.   : 10.20   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :22.99
##  1st Qu.: 77.30   1st Qu.: 0.6207   1st Qu.: 0.2542   1st Qu.: 0.2952   1st Qu.:47.76
##  Median : 90.06   Median : 2.2476   Median : 0.5498   Median : 0.8262   Median :51.67
##  Mean   : 83.65   Mean   : 9.1080   Mean   : 1.2540   Mean   : 1.9835   Mean   :51.24
##  3rd Qu.: 95.45   3rd Qu.:10.5097   3rd Qu.: 1.2210   3rd Qu.: 2.1780   3rd Qu.:55.40
##  Max.   :100.00   Max.   :85.9478   Max.   :42.6194   Max.   :41.9303   Max.   :78.08
##
##    BirthRate        deathRate
##  Min.   : 0.000   Min.   : 59.7
##  1st Qu.: 4.521   1st Qu.:161.2
```

```
##  Median : 5.381   Median :178.1
##  Mean   : 5.640   Mean   :178.7
##  3rd Qu.: 6.494   3rd Qu.:195.2
##  Max.   :21.326   Max.   :362.8
##
```

One methodology to check data integrity is to plot observations by index so that we can view all the values of a particular variable in a scatter plot to look for potentially strange patterns of data. This approach occasionally highlights errors in a data set that may be due to mistakes in data gathering/recording, malicious intervention by instructors, or otherwise.

Given the limited space in this report, we provide only plots that look suspicious (i.e., data that appear to have some pattern against Geography). The remaining plots look fine to us (i.e., data that appear random across Geography). The following plot for *avgAnnCount* variable looks suspicious. *avgAnnCount* is the only variable that we found suspicious based on the plotting approach.

```
# (ii) Data integrity check by plotting for all numeric and integer variables against the
# index, which appears to be clustered based on state, but should give a good display of
# the values on the y axis.

# "avgAnnCount" variable looks suspicious. Based on iteration of histogram plots and
# sample subset of the suspicious value, we narrowed down the culprit value to be 1962.668.
plot(dataset$X, dataset$avgAnnCount, col = ifelse(dataset$avgAnnCount > 1962.667 &
                                                  dataset$avgAnnCount < 1962.669,
                                                  'red','dark blue'),
     cex = .4,
     main = "Variable avgAnnCount by Index",
     xlab = "Index",
     ylab = "Value of avgAnnCount")
```



Based on the above approaches, our findings are as follows:

1. The *avgAnnCount* has a suspicious pattern, as shown in the plot above, in which 206 counties have the same mean incidence of 1962.668.

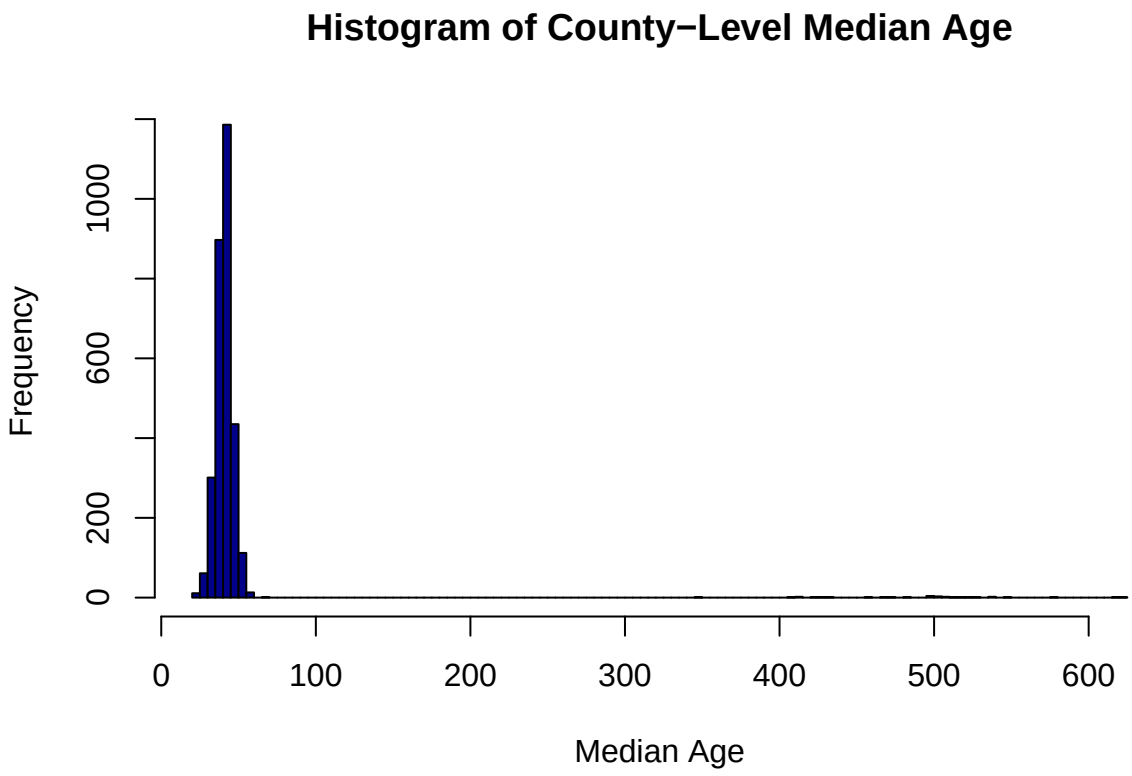
```
# Extract subset dataset that contain suspicious "avgAnnCount" value.
suspicious_incidence_subset <- dataset[dataset$avgAnnCount > 1962.667 &
                                       dataset$avgAnnCount < 1962.669, ]

# Count number of records that are subject to this potential issue.
nrow(suspicious_incidence_subset)
```

```
## [1] 206
```

For suspicious *avgAnnCount* value, we will recode it with NA. This occurs in the in the data processing and preparation section below.

2. The *deathRate* which is the key dependent variable of this study seems reasonable and doesn't contain any missing values.
3. The following Percentage and Rate variables are always complete and take values between 0 and 100: *BirthRate*, *povertyPercent*, *PercentMarried*, *PctPrivateCoverage*, *PctEmpPrivCoverage*, *PctPublicCoverage*, and *PctMarriedHouseholds*
4. Likewise, income, age, population and household size variables seem to largely take on reasonable values and don't contain missing values. These include: *medIncome*, *MedianAgeMale*, *MedianAgeFemale*, *popEst2015*, and *AvgHouseholdSize*
5. *MedianAge*, however, does seem to have a large outlier, with a maximum value of 624. We investigate further using a histogram to check how many observations are affected:

```
hist(dataset$MedianAge, main = "Histogram of County-Level Median Age",
     xlab = "Median Age", col = "dark blue", breaks = 100)
```



Looking at the values directly, perhaps these had the decimal *"accidentally"* removed. We have no reasons

other than intuition to support this. However, to avoid making an assumption, it may be best to recode the offending *MedianAge* value with NA. Again, this occurs in the next section.

6. Percentages by races should sum close to 100%, even if there is no obvious flaw in the summary statistics of each.

```
# Percentage by races should sum up close to 100%.
total_race <- dataset$PctWhite + dataset$PctBlack + dataset$PctAsian + dataset$PctOtherRace
summary(total_race)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.23   96.41   97.70   95.99   98.42  100.00
```

Based on the summary statistics above, it is unusual that the sum of all races can be as low as 11.23%. Inspecting the data further, we found that it is possible that the percentage data in some counties were recorded as decimals. For example, in row 106, it is possible that percentage of Black is 10.07% as opposed to 0.1007%; likewise, percentage for Asian could be 56.15% as opposed to 0.5615%. Another alternative explanation is that there may be an "unspecified race" that might have been omitted from the data set by error or accident. Also, there is no obvious way to elect "mixed-race" heritage.

Whatever the reason, we can correct that by adding a new variable as "unspecified race" to fill in the missing portion. For example, if the current total race is 11%, then the "unspecified race" should be 89%.

```
# Sample 3 rows that contain the issue
head(subset(dataset[, c(10, 24:27)],total_race<50), 3)
```

```
##                       Geography PctWhite  PctBlack   PctAsian PctOtherRace
## 106   Thurston County, Nebraska 41.41952 0.1007774 0.56147423    0.3887129
## 176 McKinley County, New Mexico 16.93289 0.8148869 0.96894511    3.7352361
## 456 Benson County, North Dakota 42.71416 0.0000000 0.07359435    0.6476303
```

7. Likewise, we checked if there is any obvious flaw in the education data by age group (i.e., 18-24 and above 25). We performed the following analyses, and found no obvious error, albeit some missing data that were observed. The maximum value of 100.1 which is above 100 is likely due to a minor rounding error.

```
# Percentage by education for each age group should not add up over 100%
total_edu_18_24 <- dataset$PctNoHS18_24 + dataset$PctHS18_24 + dataset$PctBachDeg18_24 +
  dataset$PctSomeCol18_24
summary(total_edu_18_24)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    99.9   100.0   100.0   100.0   100.0   100.1    2285
```

```
total_edu_25 <- dataset$PctHS25_Over + dataset$PctBachDeg25_Over
summary(total_edu_25)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.80   45.10   48.20   48.09   51.40   62.80
```

8. Percentage of employment data (i.e., *PctEmployed16_Over* and *PctUnemployed16_Over*) should not add up over 100%. Per the summary statistics below, the employment data seems to match our expectation.

```
# Sum of employed and unemployed percentages should not add up over 100%
total_employ_16 <- dataset$PctEmployed16_Over + dataset$PctUnemployed16_Over
summary(total_employ_16)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   22.40   58.00   62.50   62.01   66.60   82.70     152
```

10. Running a data integrity check for *binnedInc*, per the summary below, there are no obvious data flaws. The variable *binnedInc* is of Factor data type, which we treat it as a categorical variable.

```
# data integrity check specific to binnedInc
summary(dataset$binnedInc)
```

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

```
# check the levels of this factor data
levels(dataset$binnedInc)
```

```
##  [1] "(34218.1, 37413.8]" "(37413.8, 40362.7]" "(40362.7, 42724.4]" "(42724.4, 45201]"
##  [5] "(45201, 48021.6]"   "(48021.6, 51046.4]" "(51046.4, 54545.6]" "(54545.6, 61494.5]"
##  [9] "(61494.5, 125635]"  "[22640, 34218.1]"
```

11. We note that there are only 3007 official counties in the US and 3,242 total "county equivalents" (https://www2.census.gov/geo/pdfs/reference/GARM/Ch4GARM.pdf), so it is possible that the data set contains duplicates, includes "equivalents" or is otherwise not matching expectations.

```
# count the number of duplicated values
anyDuplicated(dataset$Geography)
```

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

We find that no Geography factor levels labels are duplicated. Perhaps some "equivalents" appear in the labels? We will search for counties that might related to Guam, Virgin Islands, and Puerto Rico

```
# see if any levels include the following text:
grep("Guam ", levels(dataset$Geography))
```

```
## integer(0)
```

```
grep("Virgin ", levels(dataset$Geography))
```

```
## integer(0)
```

```
grep("Puerto ", levels(dataset$Geography))
```

```
## integer(0)
```

A result of 0 means that no labels matched expected "equivalent" geographies. A more detailed exploration would involve standardizing county names and comparing to a known list from the US Census. This is outside of the scope of this project.

Explain whether any data processing or preparation is required for your data set.

As discussed in the previous section, we do have a couple data processing and preparation steps to take. They are detailed and addressed as follows:

1. We replace the offending *MedianAge* values and suspicious *avgAnnCount* values with NA.

```
# recode offending MedianAge value with NA
dataset$MedianAge[dataset$MedianAge > 100] = NA

# recode suspicious avgAnnCount value with NA
dataset$avgAnnCount[dataset$avgAnnCount > 1962.667 & dataset$avgAnnCount < 1962.669] = NA
```

2. The income category (as represented by the factor variable *binedInc*) may be of interest in the analysis. For example, the category "[22640, 34218.1]" should be the first rather than the last. We can easily adjust this order.

```
dataset$binnedInc <- factor(dataset$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]")
                            )
```

3. Add "unspecified race" percentage as a new data column.

```
# assign missing percentage to new variable *PctUnspecifiedRace*
dataset$PctUnspecifiedRace <- 100 - dataset$PctWhite - dataset$PctBlack -
  dataset$PctAsian - dataset$PctOtherRace

# confirm if the new variable looks fine
summary(dataset$PctUnspecifiedRace)
```

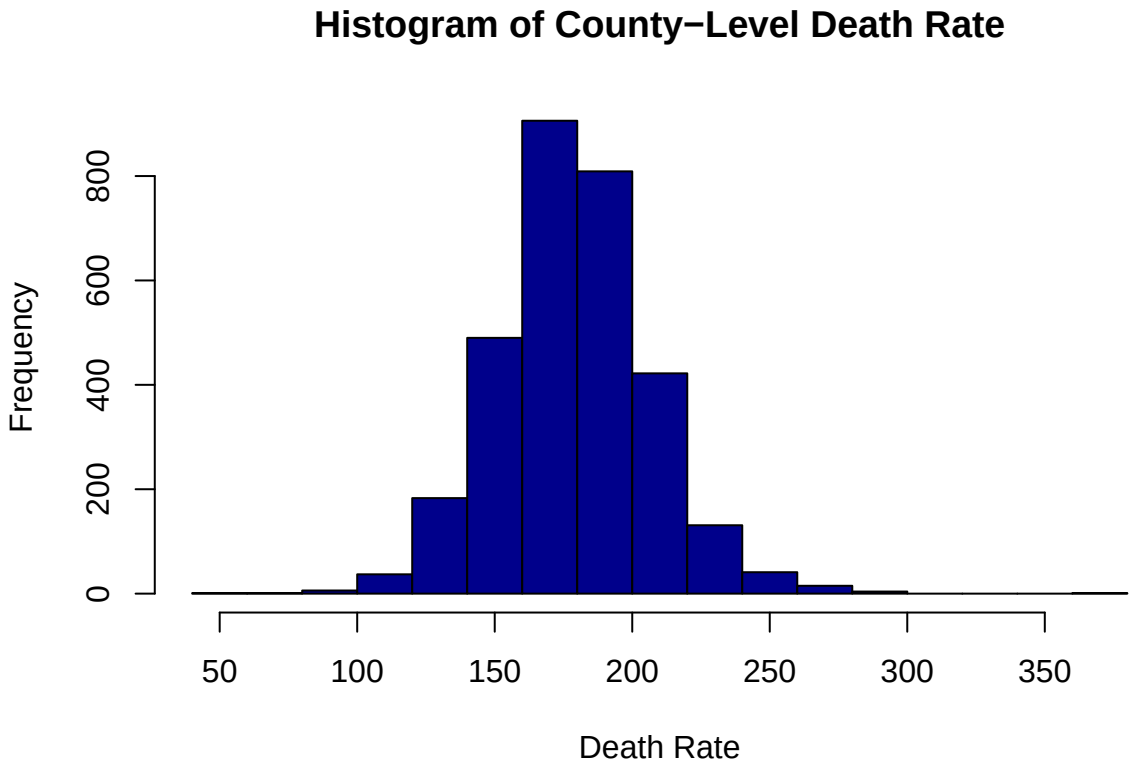
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.576   2.301   4.009   3.590  88.775
```

Univariate Analysis of Key Variables (20 pts)

We are focusing on the different types of health insurance coverage and the impacts of that coverage on the death rate. There are three variables in the data set that capture percentages of the population with a particular type of insurance coverage, these insurance types are private insurance, private employer-provided insurance, and public insurance. We plan to focus on these variables as our key variables to support our research question. Another key variable we are including is median income because income often dictates eligibility for public health insurance under the age of 65. The section below displays histograms for the death rate, as well as the insurance-related variables and the median income.

Use visualizations and descriptive statistics to perform a univariate analysis of each key variable.

```
# create histograms and summary statistics for the target variable
hist(dataset$deathRate, main = "Histogram of County-Level Death Rate",
     xlab = "Death Rate", col = "dark blue")
```

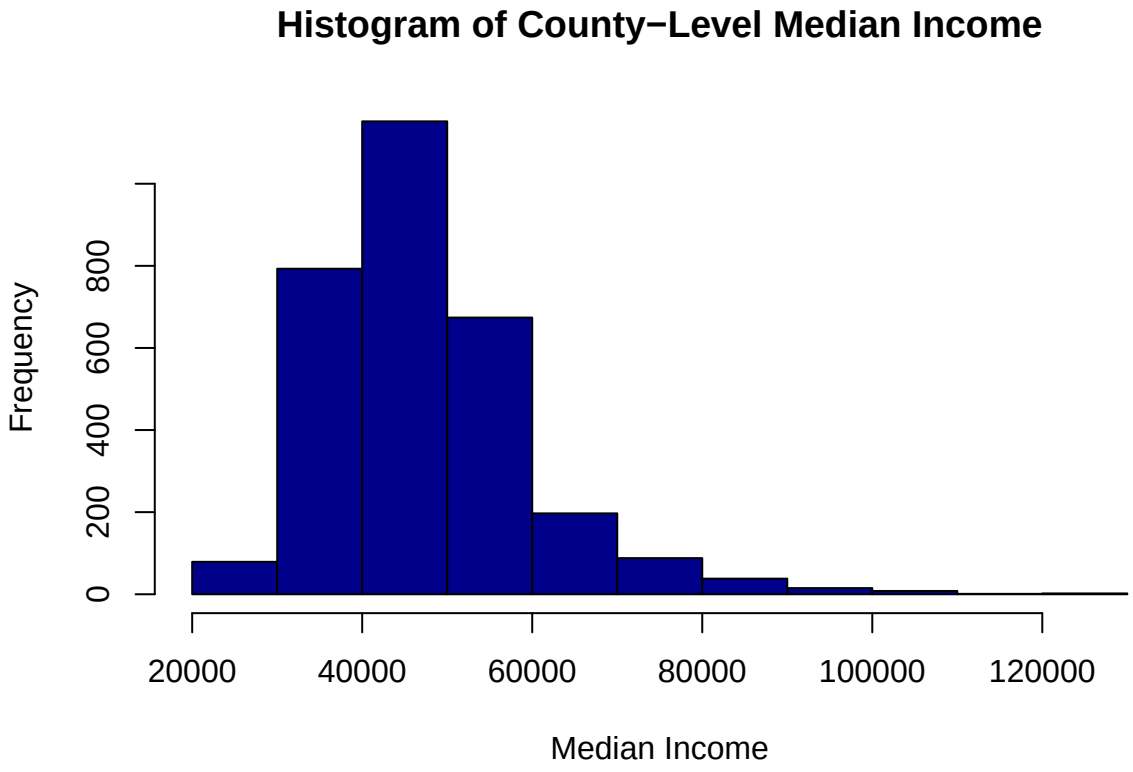



```
summary(dataset$deathRate)
```

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

The death rate is pretty evenly distributed around the mean. The maximum and minimum values that seem reasonable, however there is a pretty wide spread, from a minimum of 59.7 to a maximum of 362.8, which indicates there are pretty wide disparities between cancer death outcomes in different communities across the US.

```
# create histograms and summary statistics for the median income
hist(dataset$medIncome, main = "Histogram of County-Level Median Income",
     xlab = "Median Income", col = "dark blue")
```

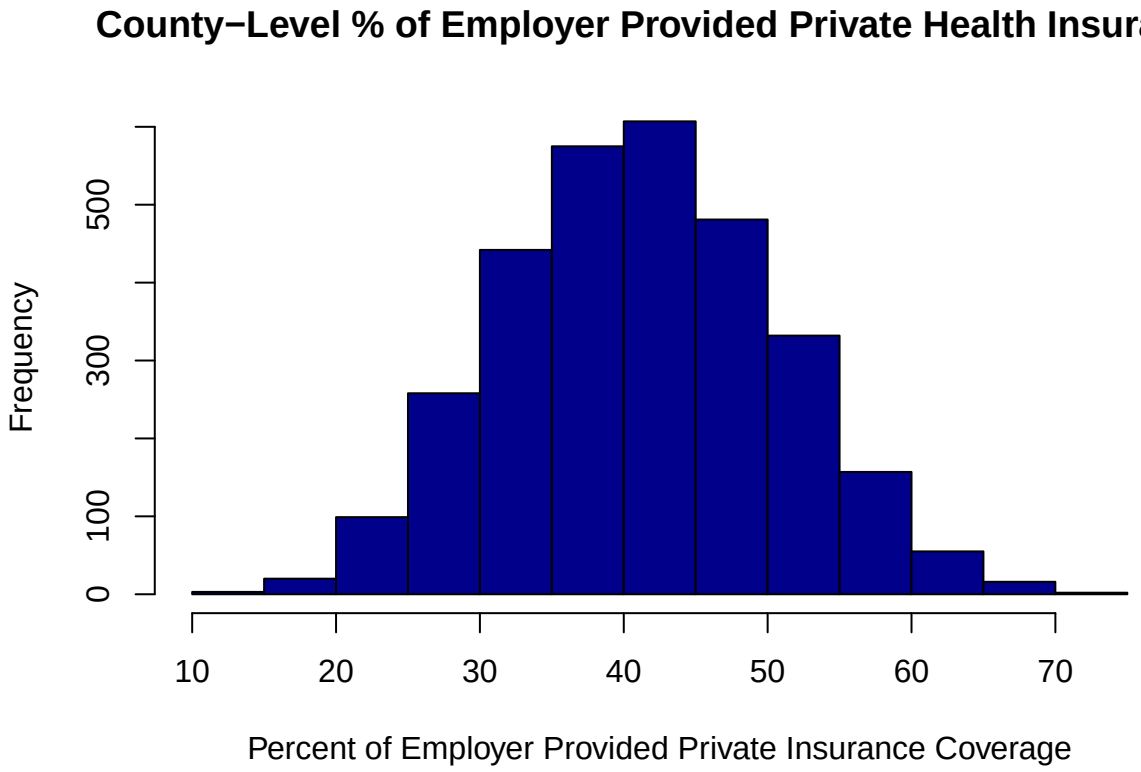



```
summary(dataset$medIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22640   38882   45207   47063   52492  125635
```

Median income appears as expected, with some counties having very low incomes in the $20,000 to $30,000 range, but most are clustered around $30,000 to $60,000. There's a pretty log tail reaching out to a maximum of just over $125,000.

```
# create histograms and summary statistics for the percent of employer-provided
# private health insurance
hist(dataset$PctEmpPrivCoverage,
     main = "County-Level % of Employer Provided Private Health Insurance",
     xlab = "Percent of Employer Provided Private Insurance Coverage",
     col = "dark blue")
```

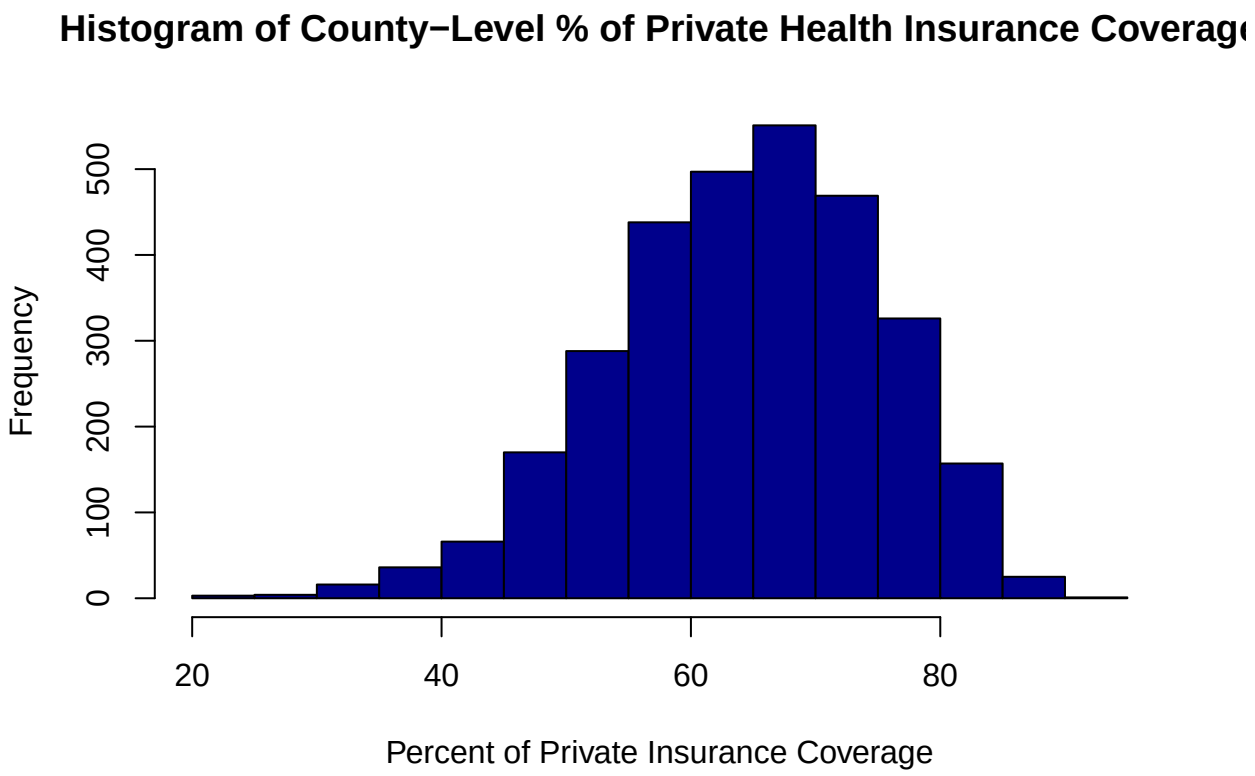



```
summary(dataset$PctEmpPrivCoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    13.5    34.5    41.1    41.2    47.7    70.7
```

The distribution for percent of employer-provided private health insurance appears to be quite normal in shape. The values range from 13.5% to 70.7% coverage, so again there are some stark disparities between the minimum and maximum values, but that does not seem unexpected.

```
# create histograms and summary statistics for the percent of private health insurance
hist(dataset$PctPrivateCoverage,
     main = "Histogram of County-Level % of Private Health Insurance Coverage",
     xlab = "Percent of Private Insurance Coverage",
     col = "dark blue")
```

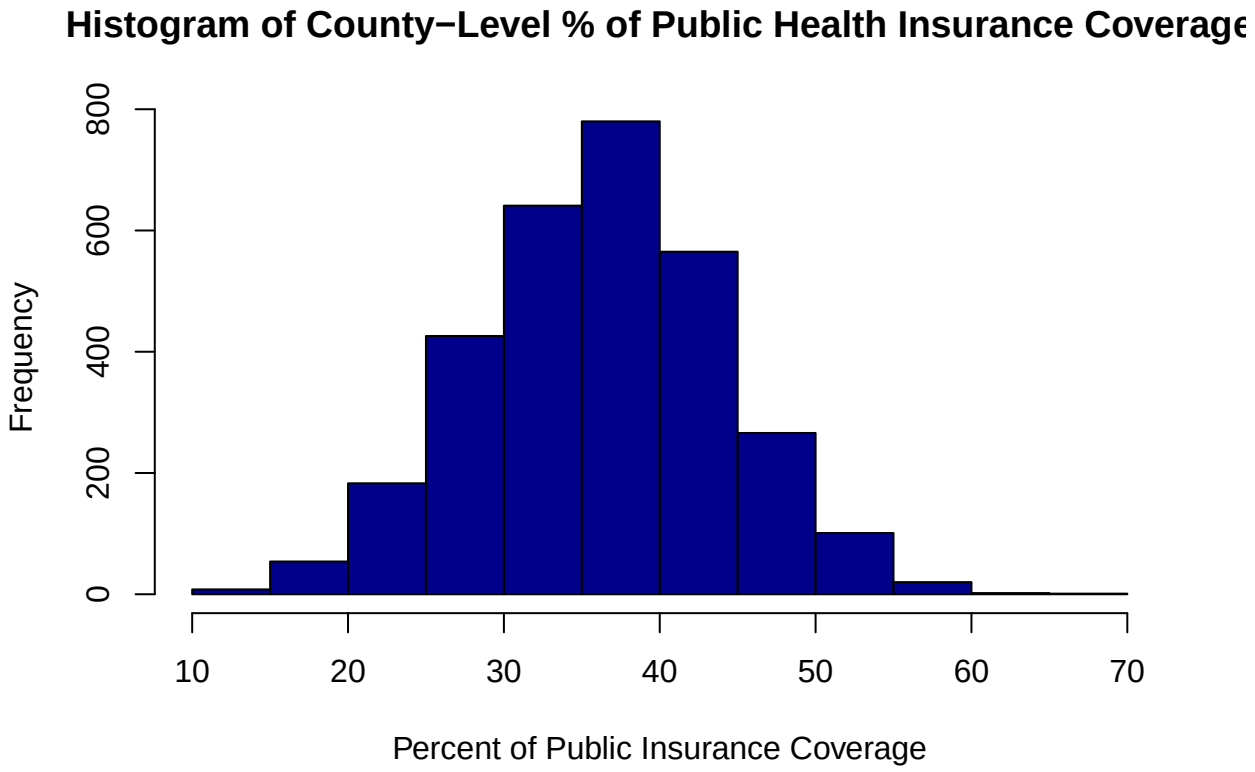



```
summary(dataset$PctPrivateCoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.30   57.20   65.10   64.35   72.10   92.30
```

The values for the percent of private health insurance coverage are within expected ranges for a percentage-based variable. The skew here is more to the left than employer-provided coverage, indicating that higher percentages of people have private health insurance coverage than have employer-provided insurance coverage, which could be due to the implementation of the Affordable Care Act.

```
hist(dataset$PctPublicCoverage,
     main = "Histogram of County-Level % of Public Health Insurance Coverage",
     xlab = "Percent of Public Insurance Coverage",
     col = "dark blue")
```




```
summary(dataset$PctPublicCoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.20   30.90   36.30   36.25   41.55   65.10
```

The distribution for percent of public health insurance coverage is also pretty evenly distributed around the mean, with ranges of values that appropriately represent a percentage variable. The rates of public health insurance coverage are lower than private and employer-provided private insurance coverage.

Be sure to describe any anomalies, coding issues, or potentially erroneous values. Explain how you respond to each issue you identify.

Analyzing these variables further, we looked to see if we could calculate the percentage of uninsured people from the percentages of insured people. Summing all three insurance-related percentages gives us only 18 values that are less than 100% and the first quartile is 132.5%, so clearly that is not an accurate interpretation of these variables.

```
# calculate sums
dataset$ins_pct_sum_all <- dataset$PctPrivateCoverage + dataset$PctEmpPrivCoverage +
  dataset$PctPublicCoverage
summary(dataset$ins_pct_sum_all)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    82.8   132.4   142.5   141.8   152.3   177.5
```

```
# number of rows that summed to less than 100%
sum(dataset$ins_pct_sum_all < 100)
```

```
## [1] 19
```

Looking to see if employer-provided private insurance may be a subset of private insurance, we check to see if the values for the percent of employer-provided private insurance is less than the percent of private insurance. Indeed the percent of private insurance is always larger than the percent of employer-provided

private insurance, so it could be a subset, but we cannot tell that for sure based solely on this analysis.

```
# number of rows where the percent of employer-provided insurance is less than the
# percent of private insurance
sum(dataset$PctEmpPrivCoverage < dataset$PctPrivateCoverage)
```

```
## [1] 3047
```

```
nrow(dataset)
```

```
## [1] 3047
```

We see in the code below that the sum of the percent of private insurance and public insurance still sums to be more than 100% most of the time.

```
# calculate sums
dataset$ins_pct_sum_pvt_pub <- dataset$PctPrivateCoverage + dataset$PctPublicCoverage
summary(dataset$ins_pct_sum_pvt_pub)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   65.40   96.25  101.30  100.61  105.80  131.70
```

```
# number of rows where the percent of employer-provided insurance is less than the
# percent of private insurance
sum(dataset$ins_pct_sum_pvt_pub < 100)
```

```
## [1] 1313
```

Each of the individual percentages seem to be within reasonable ranges, so the issue of the percentages not summing appropriately may be due to people that have more than one kind of insurance, or another unexpected caveat in the calculations.

Note any features that appear relevant to statistical analysis. Discuss what transformations may be appropriate for each variable.

All features discussed here are conceptually relevant including Age, Income, and Insurance Percentage Types. It is too early to tell if other variables like racial distributions are important to address. We also have not yet compared any variable against Cancer Mortality rates, so it is unclear which independent variables are associated with our dependent variable.

Finally, specific transformations are discussed in the next section.

Analysis of Key Relationships (30 pts)

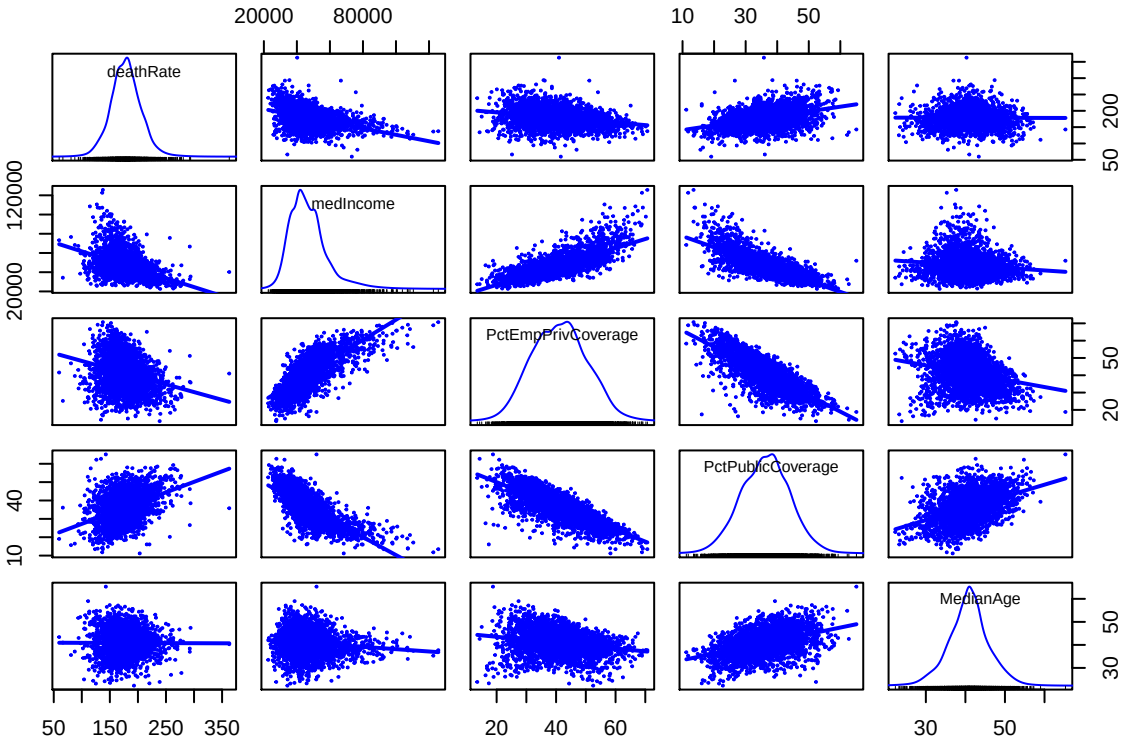
Explore how your outcome variable is related to the other variables in your dataset. Make sure to use visualizations to understand the nature of each bivariate relationship.

We use scatter plot matrices to explore relationships between selected variables. We found income and insurance features to be relevant to our statistical analysis. On the other hands, age doesn't appear to hold interesting relationship here.

```
# load library "car" for functions used below
library(car)
```

```
## Loading required package: carData
```

```
# create scatter plots and histograms of all chosen key variables
scatterplotMatrix( ~ deathRate + medIncome + PctEmpPrivCoverage + PctPublicCoverage +
                     MedianAge, data = dataset , smooth= FALSE, cex= .2)
```



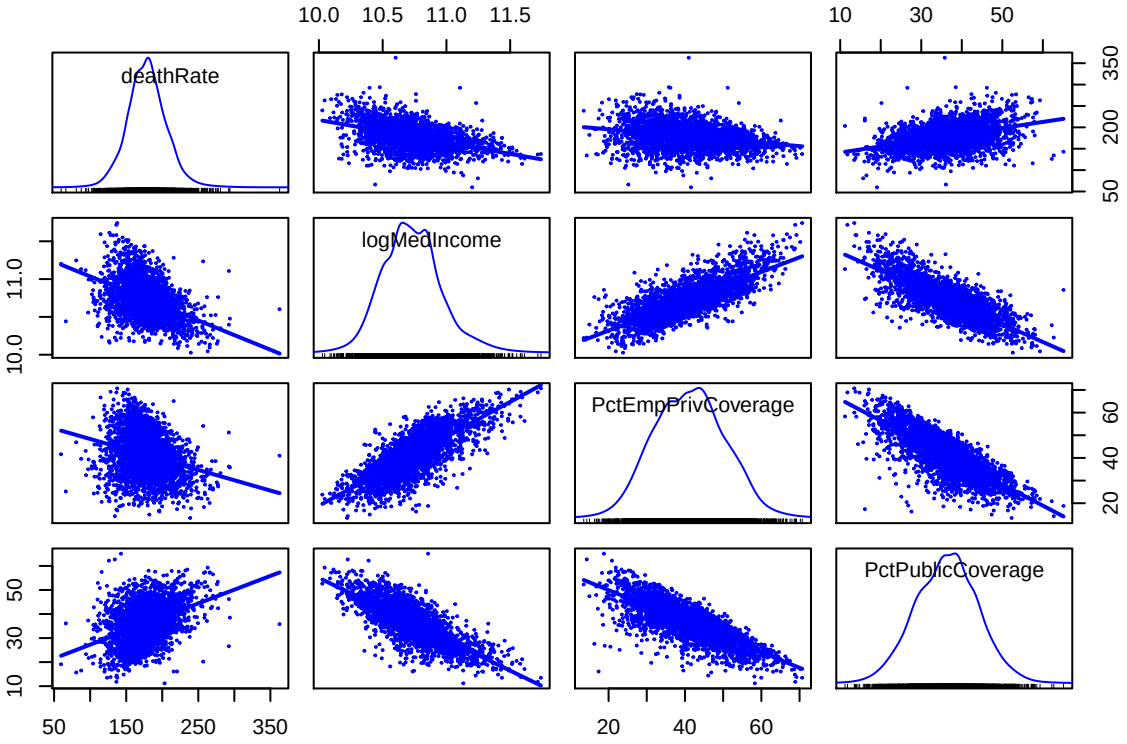

```
# list correlations in a grid
#library(corrplot)
#corrplot.mixed(cor(dataset[,c(30,3,22,23,7)], use = "complete.obs"))
```

Before we jump to discussion of bivariate relationships, we note that *medIncome* isn't symmetrically distributed like the other variables.

What transformations can you apply to clarify the relationships you see in the data? Be sure to justify each transformation you use.

We apply a log transformation to *medIncome* in order to spread out variation occurring on the low ends of scales. This allows us to visually clarify its relationship with other variables, use methods for linear associations, and discuss associations on a percentage-basis rather than an absolute-basis. We then re-run the visualization.

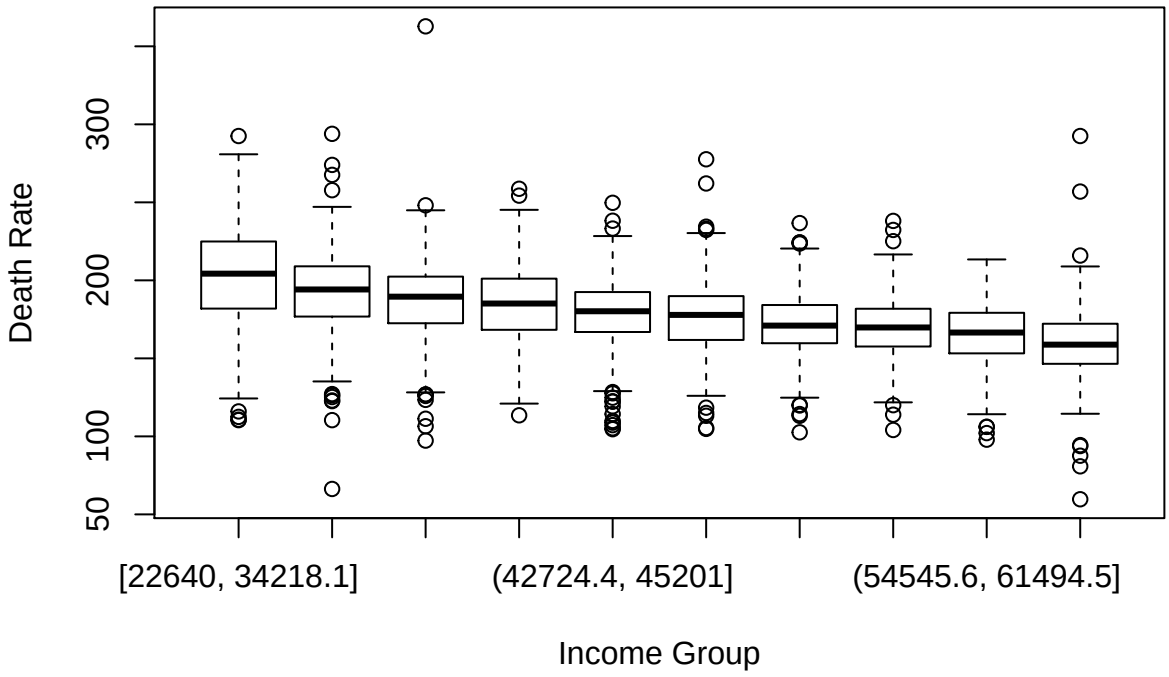
```r
# log transform to medIncome variable
dataset$logMedIncome <- log(dataset$medIncome)

# create scatter plots and histograms of all chosen variables
scatterplotMatrix( ~ deathRate + logMedIncome + PctEmpPrivCoverage + PctPublicCoverage,
                   data = dataset , smooth= FALSE, cex= .2)
```




```r
# verify that the linear correlation has improved post transformation
pre_log_cor  <- cor(dataset$deathRate, dataset$medIncome)
post_log_cor <- cor(dataset$deathRate, dataset$logMedIncome)
abs(post_log_cor) > abs(pre_log_cor)
```

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

```r
# create box plots to visualize relationship between income category and death rates
boxplot(deathRate ~ binnedInc, data = dataset,
        main = "Death Rate by Income Group",
        xlab = "Income Group", ylab = "Death Rate")
```



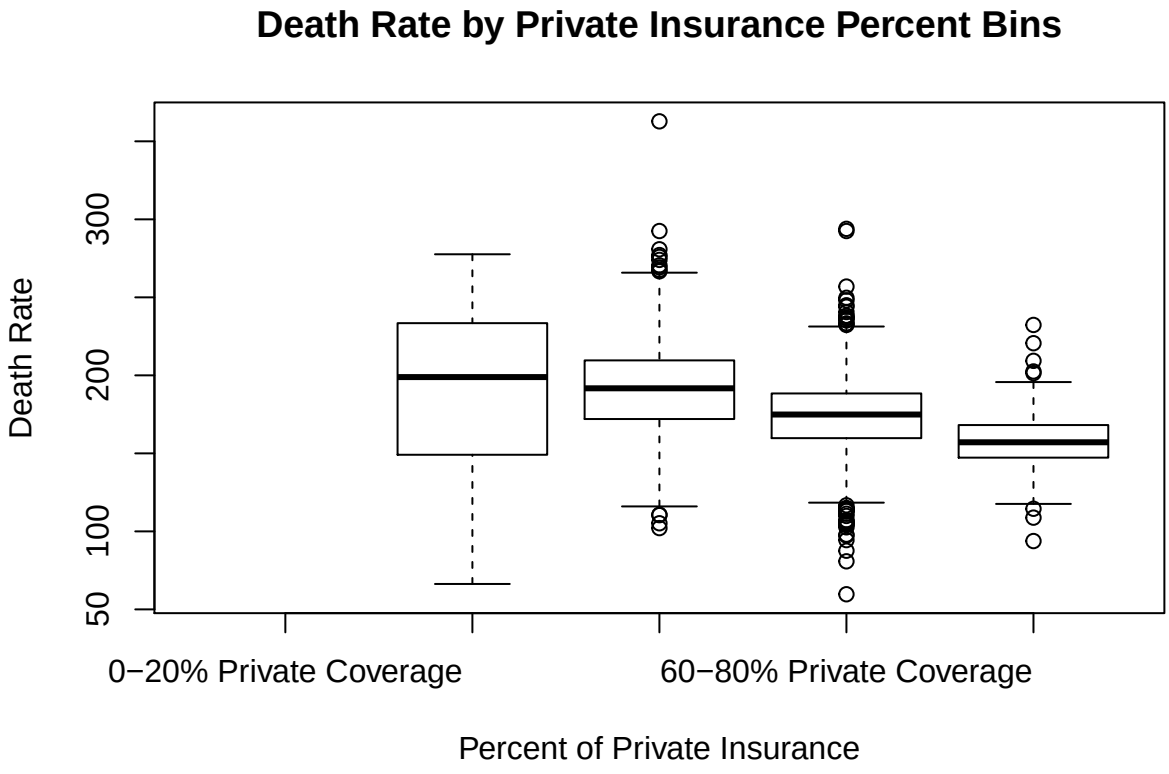
We find that Death Rates (*deathRate*) are inversely correlated with Income (*logMedIncome*), inversely correlated with Employer Provided Coverage Rate (*PctEmpPrivCoverage*), and positively correlated with Public Coverage Rate (*PctPublicCoverage*).

Additionally, we see some expectation-confirming relationships between Income, Employer Provided Coverage, and Public Coverage:

- logMedIncome is positively correlated with Employer Provided Coverage. This is to be expected, since higher-compensated occupations are more likely to provide coverage to employees.
- logMedIncome is inversely correlated with Public Coverage, using the reverse of the logic above.
- Private and Public coverage are inversely correlated, since they are both covering some mutually-exclusive portion of a total 100% of the population.

Now we look in more details at death rate and the relationship with different types of insurance coverage. One would expect that having health insurance increases a person's ability to afford costly cancer treatments. The following section segments the insurance coverage percentages into bins from 0% to 20%, over 20% to 40%, over 40% to 60%, over 60% to 80% and above 80%.

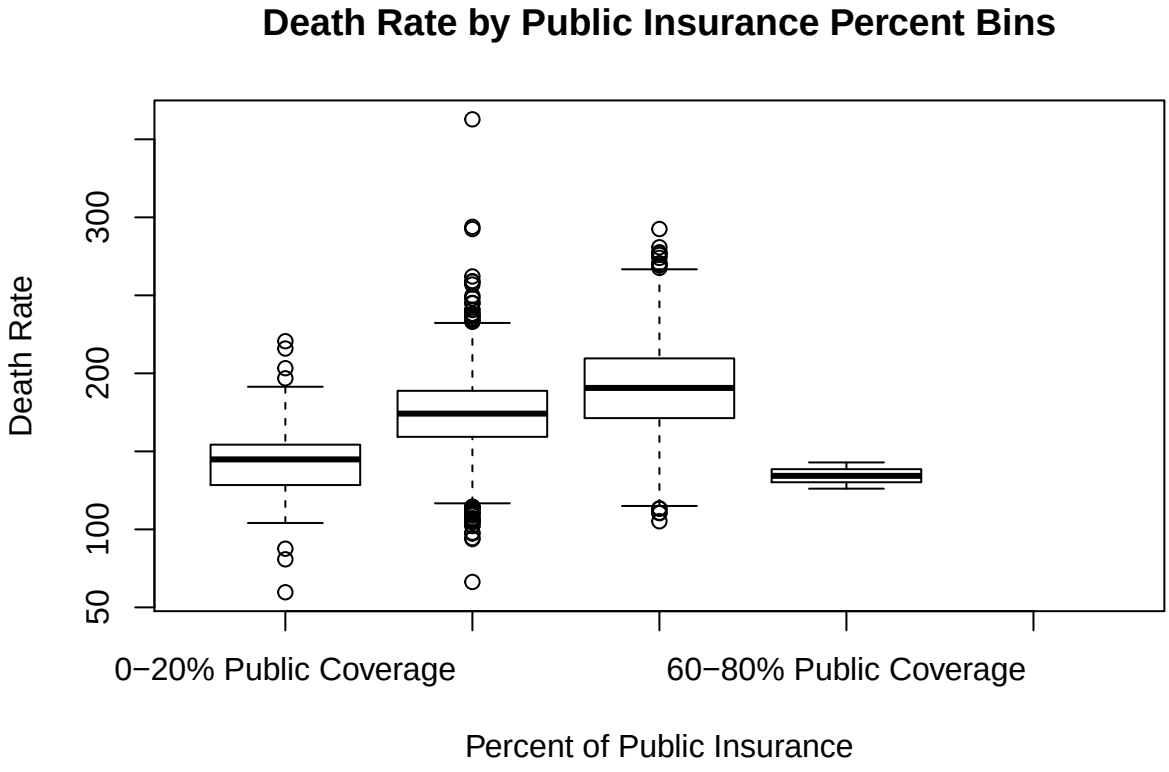
```
# create box plots to visualize relationship between private insurance rates and death rates
private_ins_bin = cut(dataset$PctPrivateCoverage, breaks =c(0,20,40,60,80, Inf),
                      labels = c("0-20% Private Coverage", "20-40% Private Coverage",
                                 "40-60% Private Coverage", "60-80% Private Coverage",
                                 "Over 80% Private Coverage"))

boxplot(deathRate ~ private_ins_bin, data = dataset,
        main = "Death Rate by Private Insurance Percent Bins",
        xlab = "Percent of Private Insurance", ylab = "Death Rate")
```



Indeed the above graph shows that communities with higher rates of private insurance, have on average a lower death rate. However, there are counties with relatively low rates of private insurance coverage that have very low death rates, and counties with 60-80% coverage that have high cancer death rates.

Now we look at public coverage rates similarly to the above analysis.

```
# create box plots to visualize relationship between public insurance rates and death rates
public_ins_bin = cut(dataset$PctPublicCoverage, breaks =c(0,20,40,60,80, Inf),
                     labels = c("0-20% Public Coverage", "20-40% Public Coverage",
                                "40-60% Public Coverage", "60-80% Public Coverage",
                                "Over 80% Public Coverage"))

boxplot(deathRate ~ public_ins_bin, data = dataset,
        main = "Death Rate by Public Insurance Percent Bins",
        xlab = "Percent of Public Insurance", ylab = "Death Rate")
```



The graph here appears to indicate that higher rates of public health insurance coverage of up to 60% actually corresponds to an increasing death rate. The death rate in the bin second to last (i.e., public overage over 60%) seems suspiciously low. Further investigation shows that there are only three counties where the percent of public health insurance coverage is above 60%, so the data in the graph above is likely only reliable up to 60% public heath insurance coverage and the downward shift seen in the graph for that segment is likely just noise given the very small sample size of three.

```
# number of observations in the last two bins
sum(dataset$PctPublicCoverage > 60)
```

```
## [1] 3
```

```
sum(dataset$PctPublicCoverage > 80)
```

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

As we saw in the scatter plot matrix above, high rate of public health insurance coverage tends to be associated with lower median incomes. Further lower median incomes correlate to higher death rates, which then might explain why increasing rates of public coverage up to 60% correlates to higher death rates.

Analysis of Secondary Effects (10 pts)

What secondary variables might have confounding effects on the relationships you have identified? Explain how these variables affect your understanding of the data.

There are other variables that we haven't explored that might have a meaningful relationship with the death rate. For example, education level, and races. In this section, we will explore both as potential secondary factors that might have a confounding effect on the death rate.

First, we study the effect of racial mix on detah rates. We first create a new variable in the dataset that will hold the value of predominant race. This variable will be categorical and will be used in boxplots to show potential relationships between racial mix with the death rate.

```
# Create a placeholder dataset that contains racial information only.

races = dataset[, c("PctWhite",
                    "PctBlack",
                    "PctUnspecifiedRace",
                    "PctAsian",
                    "PctOtherRace"
                    )
                ]

# Create a variable that records predominant race, defined as
# the race that has the highest percentage in the county.
races$predominant = names(races)[apply(races, 1, which.max)]

# Create and assign PredominantRace variable in the dataset.
dataset$PredominantRace = races$predominant

# Verify that there are asssignment to all races. The sum of the numbers below
# equal to the the total observations 3,047.
sum(dataset$PredominantRace == "PctWhite")
```

```
## [1] 2902
```

```
sum(dataset$PredominantRace == "PctBlack")
```

```
## [1] 112
```

```
sum(dataset$PredominantRace == "PctUnspecifiedRace")
```

```
## [1] 29
```

```
sum(dataset$PredominantRace == "PctAsian")
```

```
## [1] 3
```

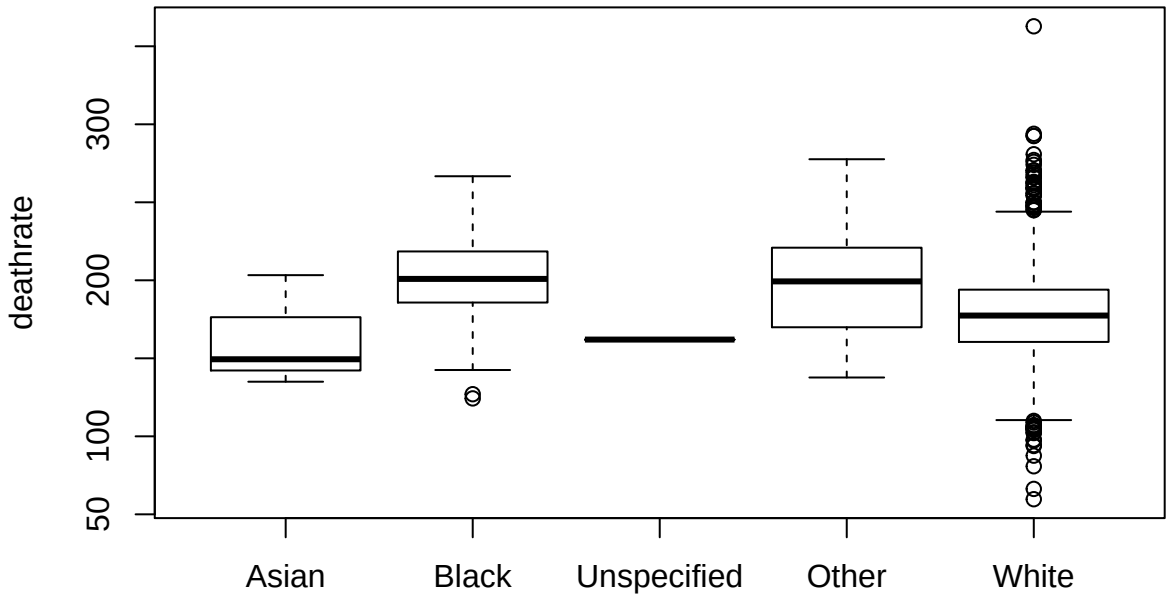
```
sum(dataset$PredominantRace == "PctOtherRace")
```

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

As shown in the boxplot below, there appears to be some relationship between races and death rates, in that counties that are predominantly Black or Unspecified race tends to have relatively higher death rates than those that are predominantly White. However, for counties that are predominantly white, the dispersion of death rates is quite large, suggesting that other factors in addition to races mix may be at play.

Note that in this analysis, we excluded predominant Asian and Other as the number of observations were too low – 3 and 1 respectively.

```r
# Box plot of death rate by predominant race
boxplot(deathRate ~ PredominantRace, data = dataset, ylab = "deathrate",
        names = c("Asian", "Black",
                  "Unspecified", "Other","White"))
```

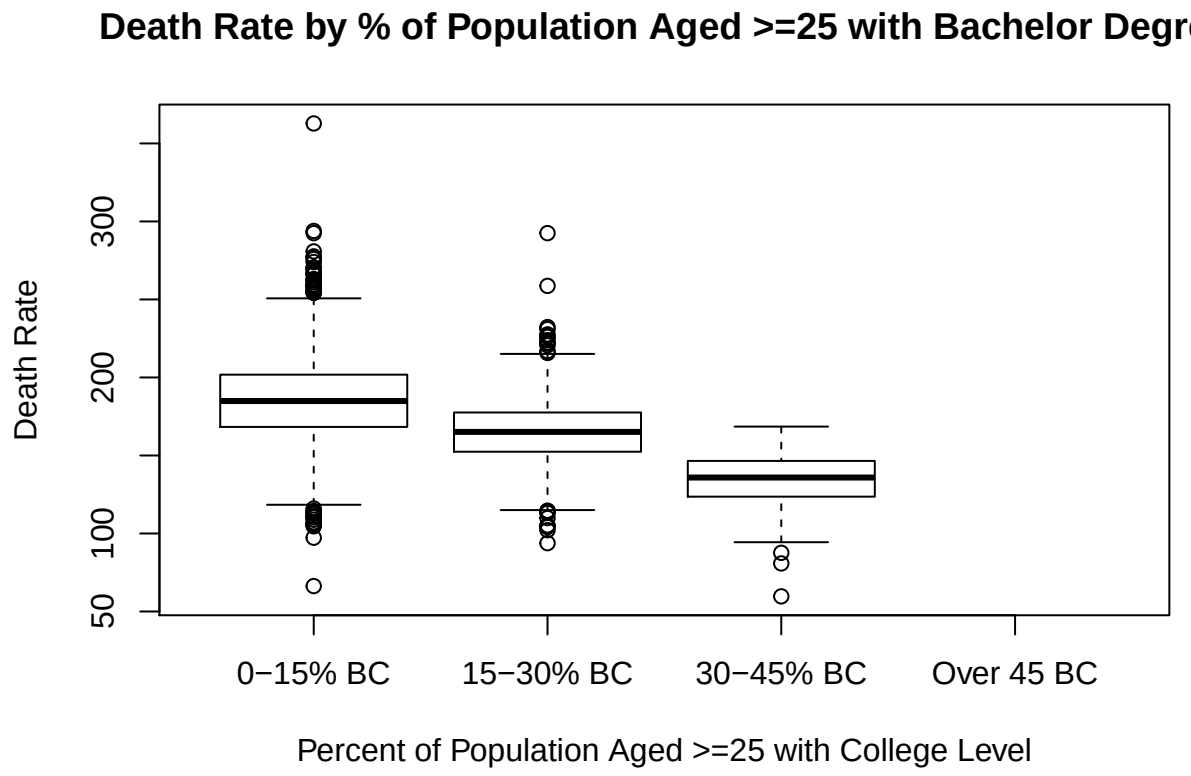


In addition, we study the effect of educational level mix on detah rates. In this study, we focus on the education level among the group of 25 years of age and older, which is split between a group that has high school degree (*PctHS25_Over* variable) and a group that has bachelor degree (*PctBachDeg25_Over* variable).

We decided to study the group of 25 years of age and older, as opposed to the group of ages between 18-24 years old, because demographically the population with age 25 years old and older has higher repsentation than those of age between 18 and 24 years old. (See https://en.wikipedia.org/wiki/Demography_of_the_United_States#Ages)

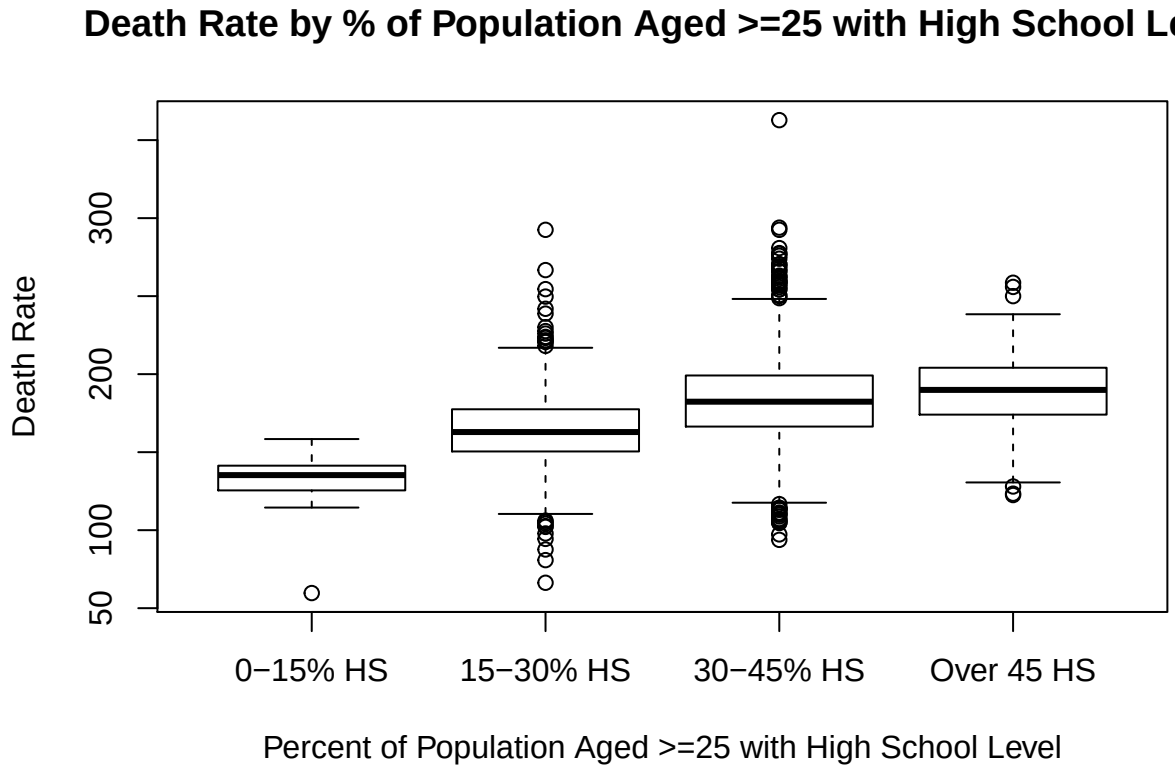
We create box plots that visualize the relationship between education levels and death rates. As shown in the the two box plots below, there are obvious trends that counties with higher proportion of population with bachelor degree tend to have lower death rates, and counties with higher proportion of population with high school degree tend to have higher death rates. Clearly there is a tendency to be drawn here that educational level may have some secondary effects on the death rates, and is worth exploring further in more details. The study of which is out of scope in this project.

```
# create box plots to visualize relationship between percent of population aged
# 25 and above with bachelor degree and death rates
education_bachelor_level = cut(dataset$PctBachDeg25_Over, breaks =c(0,15,30,45,Inf),
                      labels = c("0-15% BC", "15-30% BC",
                                 "30-45% BC", "Over 45 BC"
                                 )
                     )

boxplot(deathRate ~ education_bachelor_level, data = dataset,
        main = "Death Rate by % of Population Aged >=25 with Bachelor Degree",
        xlab = "Percent of Population Aged >=25 with College Level", ylab = "Death Rate")
```




```
# create box plots to visualize relationship between percent of population aged
# 25 and above with high school degree and death rates
education_hs_level = cut(dataset$PctHS25_Over, breaks =c(0,15,30,45,Inf),
```

```
                      labels = c("0-15% HS", "15-30% HS",
                                 "30-45% HS", "Over 45 HS"
                                 )
                      )
boxplot(deathRate ~ education_hs_level, data = dataset,
        main = "Death Rate by % of Population Aged >=25 with High School Level",
        xlab = "Percent of Population Aged >=25 with High School Level", ylab = "Death Rate")
```



Conclusion (20 pts)

Summarize your exploratory analysis. What can you conclude based on your analysis?

At the start of this analysis, we expected to see that insurance coverage would decrease death rates due to cancer because people insurance would have access to the care and treatment needed to increase survival rates. What we actually found was a relationship that is more complex than we initially thought. This analysis shows that the type of insurance people in a community have access to has different directional relationships with the death rate - higher instances of private health insurance does correspond to lower death rates; however, higher rates of public health insurance has the opposite effect. Further digging reveals that the inverse relationship between public health insurance and death rate is likely driven by the fact that people that are eligible for public health insurance are either over 65 and/or low income. It is likely that these two factors, lower income and higher median ages, both of which have positive relationships with death rate, are what is driving the negative relationship between public health insurance and death rate.

In addition to insurance type and income level, there are other factors that might explain cancer mortality. For example, the racial mix analysis in which counties were characterized by race reveales some meaningful relationship between "predominant" race and cancer mortality. Given the racial data is provided only in percentage as a characteristic of a county, we could only draw a directional conclusion that races may play a role in cancer mortality. Effects of races on cancer mortality may be a good area to study in more details.

In conclusion, it would be reasonable to recommend a formal statistical study of the effects of insurance rates, both public and private, on cancer mortality rates that controls for the effects of other health outcome drivers available in this data set, including racial mix and educational attainment. A multiple OLS analysis would therefore help to steer the Commission into making more informed recommendations on the expansion or contraction of public insurance availability or subsidy in order to improve cancer survival rates, conditioned on a better understanding of other effects.