Cancer Mortality Exploration

Zeliang (Doug) Xu, Arjun Chakraborty, John Boudreaux

An exploratory analysis of county characteristics as they relate to cancer death rates

Introduction

This exploratory analysis is meant to address the research topic:

What are the key contributing factors to high cancer death rate in some US counties? How might we be able to reduce high cancer death rates?

By looking at the dataset and our general intuition about cancer, we believe that there are two major sources of contributing factors that lead to higher cancer death rate for each physical region:

- Higher cancer occurance rates will lead to more cases of cancer deaths in a particular region
- Worse cancer treatment facilities, policies, personnel, etc. will lead to more cancer deaths in a particular region

To explore this research question and sub-topics, we will employ various statistical techniques including subsetting the data, descriptive statistics, outlier detection and treatment, multivariate correlations, and others. In particular, our analysis will focus heavily on using graphical methods to illustrate various aspects about our dataset. We find certain relationships, particularly in the area of insurance coverage, that prove interesting as levers for policy change.

It is important to note that our analysis is not meant to prove any sort of causality. All of our methods and conclusions are purely meant as descriptive methods to explore the relationships within the data. Additionally, this dataset has some very particular issues that indicate further data collection or clarification may be necessary to fully validate the results. For instance, there is no documentation on the basis of the death rate- is this a death rate per 100,000 people, averaged over some period? Is it normalized in any way? Is it only deaths from cancer? Without further documentation, we can still perform some analysis but still need to clarify some assumptions before action can be responsibly taken.

In this report, we will use R (>=3.4) to perform the analysis. We will provide code for all graphics and analysis performed so that the reader can reproduce this analysis with the given dataset.

Data Preparation

Below is a collection of libraries that we use for the study:

In [18]:
```
library(car)   # to enable csv import, plots functions
library(stringr)   # to enable string splitting
library(corrplot)   # to enable corrplot
library(data.table)   # to be able to change the indexing of data frame
options(repr.plot.width=5, repr.plot.height=4) # formatting for reasonable plo
t size
```

In [19]:
```
#may need to change address depending on where you have it saved
data <- read.csv("../project materials/cancer.csv")
```

In [20]:
```
dim(data)
```

3047 30

Our data has 3047 observations of 30 variables. These dimensions ought to be sufficient for many subsetting operations we may want to do with the data. The variables all describe different properties of US counties, primarily focusing on demographic data.

In [64]:
```
# normally would do summary()... but this does not print for good report
str(data)
```

```
'data.frame':   3047 obs. of  33 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 2084
8 13088 843954 ...
 $ povertyPercent       : num  11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 1
3.1 ...
 $ 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            : chr  "Kitsap County, Washington" "Kittitas County, W
ashington" "Klickitat County, Washington" "Lewis County, Washington" ...
 $ 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 5
6.6 ...
 $ 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 ...
 $ total_race           : num  91 96.2 94.9 95.1 95.5 ...
 $ logpopEst2015        : num  12.47 10.68 9.95 11.24 9.24 ...
 $ state                : Factor w/ 51 levels "Alabama","Alaska",..: 48 48 48
48 48 48 48 48 48 48 ...
```

From the summary statistics, boxplots, and histograms for each parameter *(most excluded from this report for the sake of brevity; can easily be reproduced using the `boxplot` and `hist` functions in R)* we found a few peculiarities in the data that we must address before going further. First, over half of our observations for `PctSomeCol18_24` are missing values. We should be wary of any sort of statistically significant results we obtain with this variable, since sampling effects will likely alter some of the true dynamics in the data. There are also missing values in `PctEmployed16_Over`, but this is a much smaller number and does not necessarily warrant the same degree of caution for `PctSomeCol18_24`.

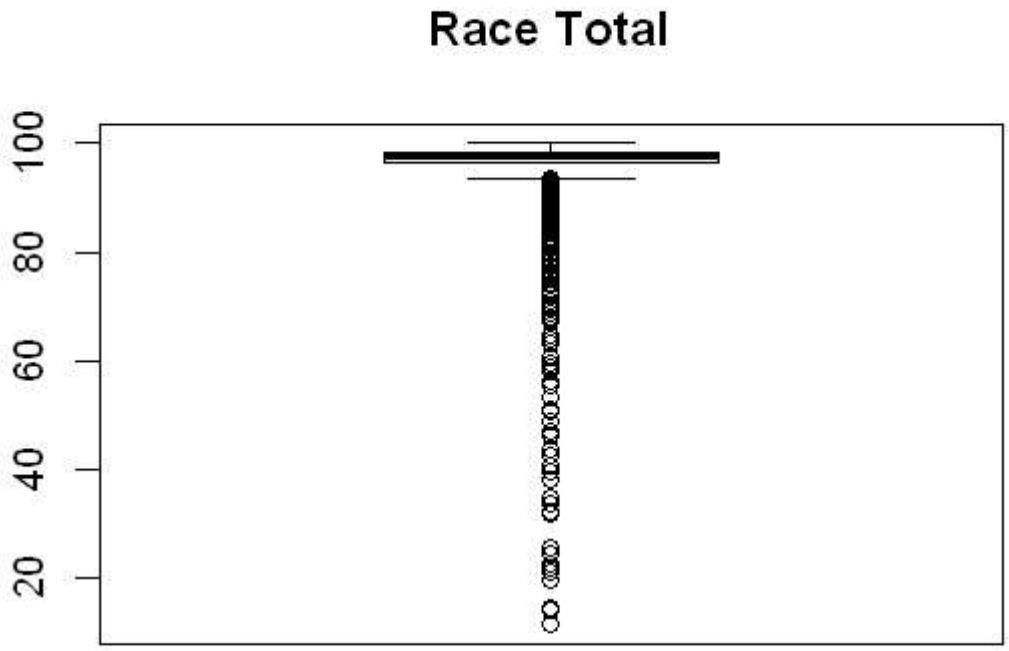
In addition to missing values, there are some values that don't make a ton of sense with reality. Looking at the columns denoting race (`PctWhite`, `PctBlack`, `PctAsian`, `PctOtherRace`) it does not seem intuitive that any county in the US would truly be 100% or 0% any value. When we check to see if the total sum of all races adds up to 100, we see that there are many examples that do not come close to 100% accounted for, with over 10% of observations not adding up to 95%. Our analysis will not focus deeply on race relationships, due to the subadequate data. Had we looked deeper into these, it may have been prudent to do some sort of data transformation to get rid of the skew in the data.

In [22]:
```
data$total_race <- data$PctAsian + data$PctBlack + data$PctWhite + data$PctOth
erRace
message("The details of data for total race percentage is: ")
summary(data$total_race)
message("There are ",nrow(subset(data,data$total_race < 95)), " rows of data w
ith total race percentage smaller than 95%")
message("There are ",nrow(subset(data,data$total_race < 50)), " rows of data w
ith total race percentage smaller than 50%")
message("There are ",nrow(subset(data,data$total_race < 20)), " rows of data w
ith total race percentage smaller than 20%")

boxplot(data$total_race, main = "Race Total")
```

```
The details of data for total race percentage is:
```

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

```
There are 421 rows of data with total race percentage smaller than 95%
There are 24 rows of data with total race percentage smaller than 50%
There are 6 rows of data with total race percentage smaller than 20%
```



When we look at the `MedianAge` column, we see some unreasonable high outliers. By examining the boxplot below, we found that values above 65 were nonsensical. These were assigned the `NA` value to be exlcuded from further analysis. This should not effect the analysis much, as these are only 31 values being removed (~1% total observations).

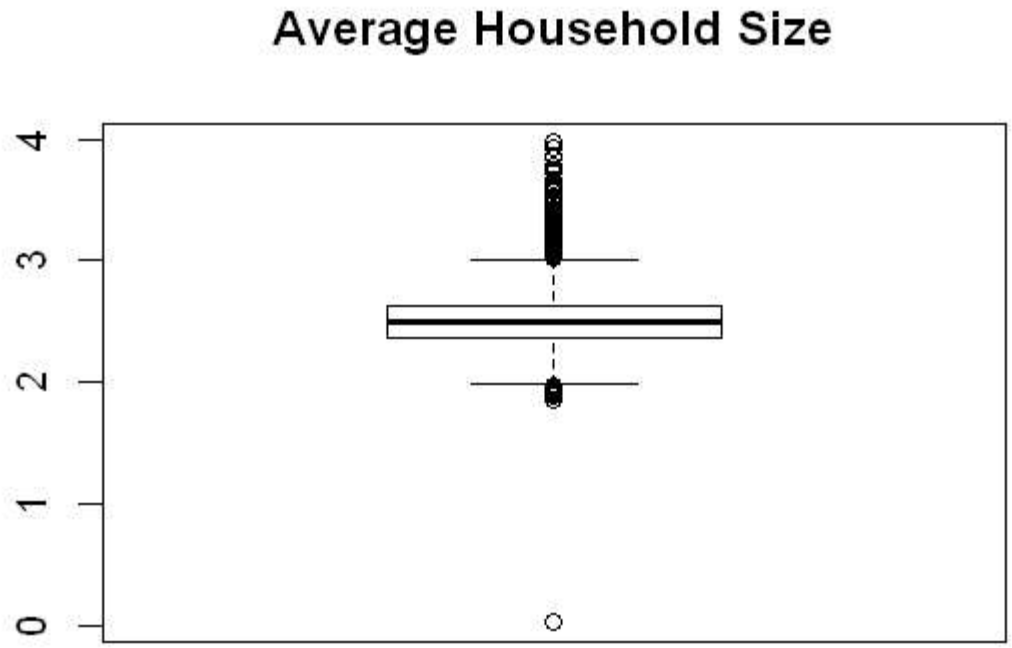
In [23]:
```
boxplot(data$MedianAge, main = "Median Age")

#filter out outliers
data$MedianAge[data$MedianAge > 65] <- NA
sum(is.na(data$MedianAge))
```

31

We also found that there are 61 instances of an Average Household Size being less than 1. While this may have some reasoning in real life such as secondary properties, there was no such documentation to confirm this. As such, we removed all values less than 1 since this does not make sense with our default assumptions that one person must live in a place for it to be considered a household. Again, this is a small amount of observations and should not effect many results in a major way.

In [24]:
```
boxplot(data$AvgHouseholdSize,
        main = "Average Household Size")

data$AvgHouseholdSize[data$AvgHouseholdSize < 1] <- NA
sum(is.na(data$AvgHouseholdSize))
```

61



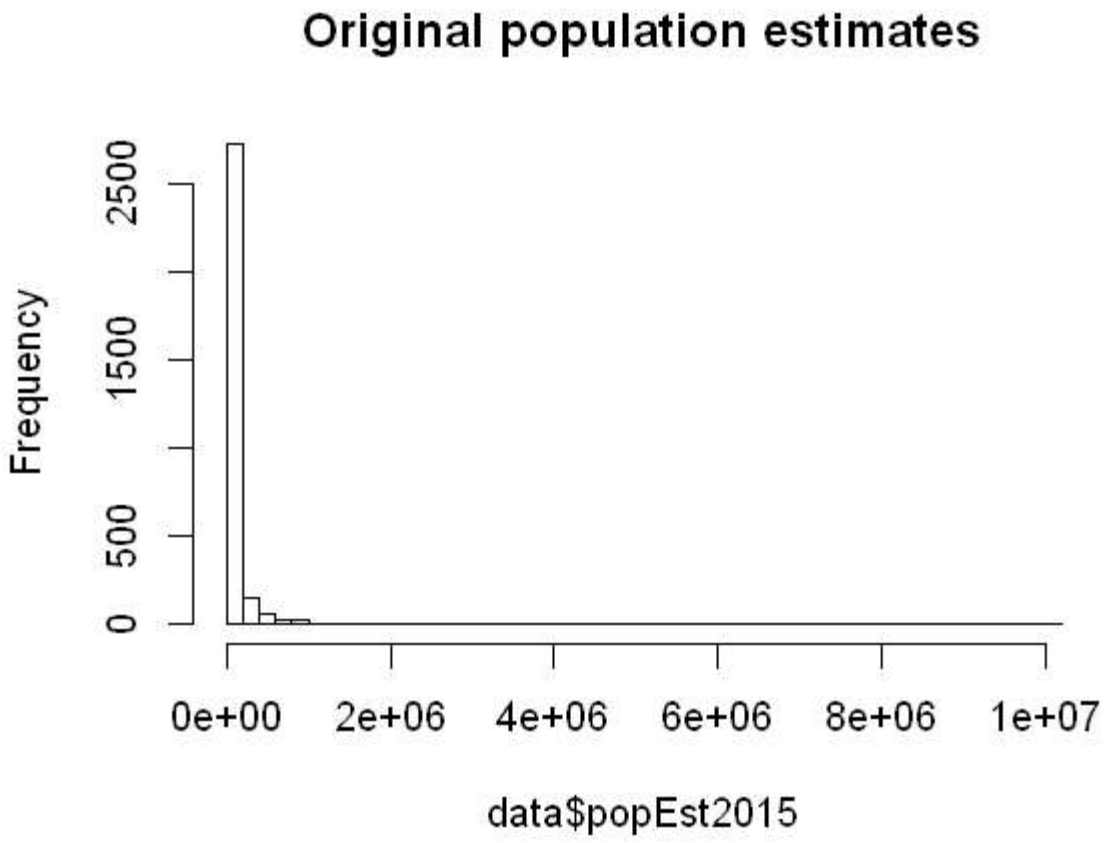
We found one value where `PctHS18_24` was zero. In the US, it is difficult to believe that not one person in any given county has completed high school. We take away this value from further analysis.

In [25]:
```
data$PctHS18_24[data$PctHS18_24 == 0] <- NA
sum(is.na(data$PctHS18_24))
```

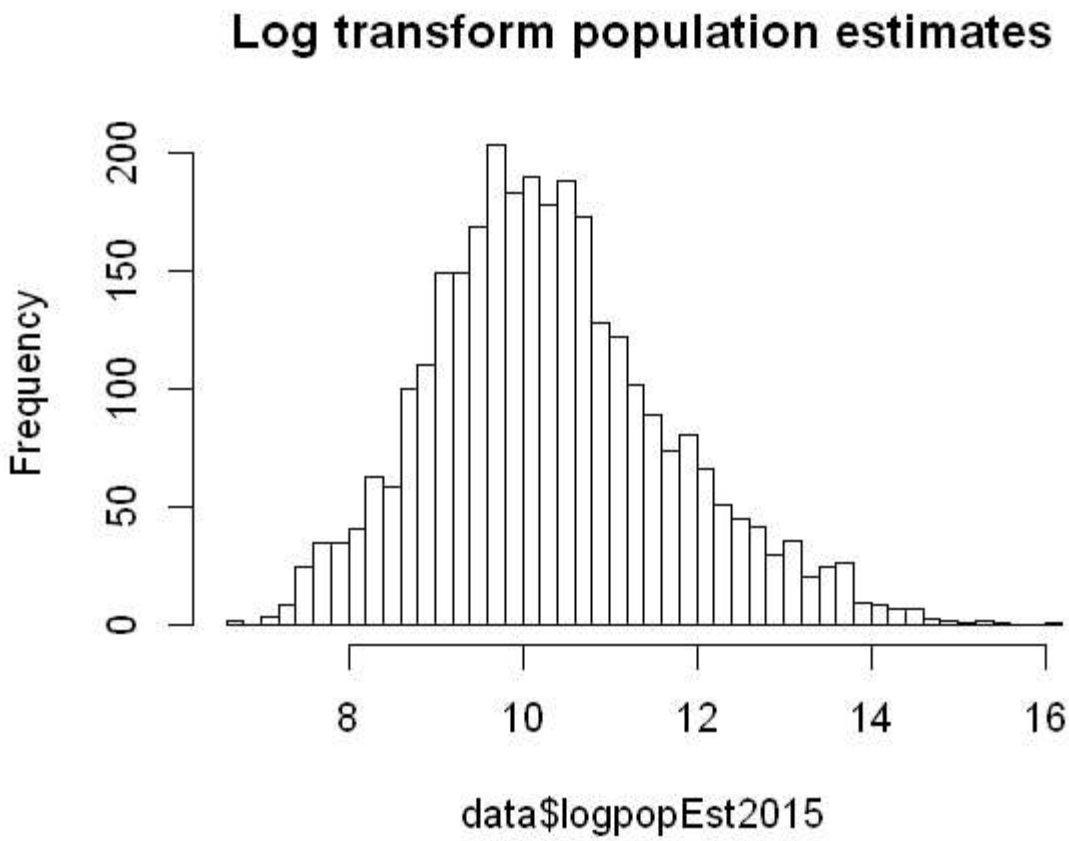
1

The population data in `popEst2015` shows very heavy skew towards lower values, but the higher values also make sense for the hyper-urbanized areas such as Los Angeles, Chicago, New York, and the others that they represent. In this case, a log transformation could prove useful for further analysis to have a more uniform distribution of values.

In [26]:
```
hist(data$popEst2015, , breaks = 50, main = "Original population estimates")
```



In [27]:
```
data$logpopEst2015 <- log(data$popEst2015)
hist(data$logpopEst2015, breaks = 50, main = "Log transform population estimat
es")
```



The last major item we have to address is the variables that are factors. The `binnedInc` column contains a categorical value for the bracket of the median income. Since this value is inherently contained by the `MedIncome` variable, we largely ignored this category since we can just as easily subset it on our own. There is another value that contains structured data in a string format in the `Geography` field, which gives the unique county for each observation. Individual counties are probably not so useful, but having the states as a factor to subset the data could prove useful.

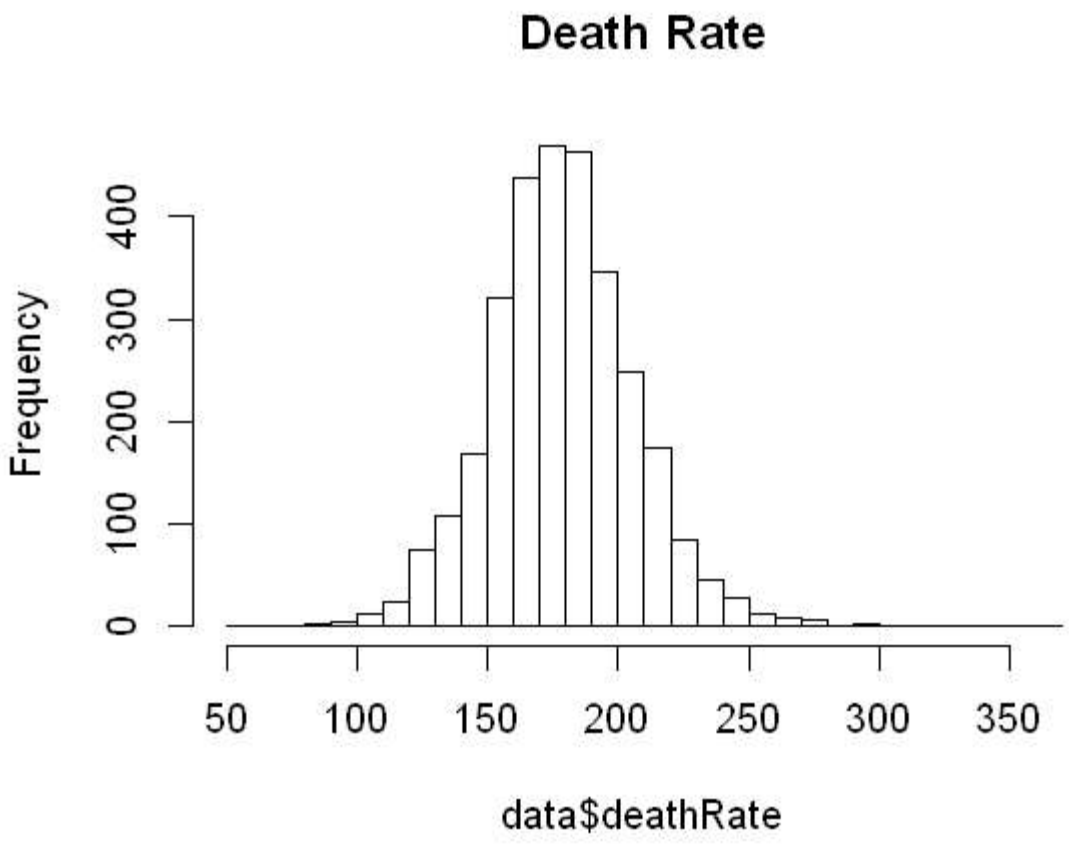
In [28]:
```
#create the state column
data$Geography <- as.character(data$Geography)
split.geo <- strsplit(as.character(data$Geography), ", ")
states <- sapply(split.geo,
                 function(x){
                   return(x[2])
                   })
data$state <- as.factor(states)
sample(data$state, 5)
```

Montana Florida Texas Texas Indiana

Exploratory Data Analysis

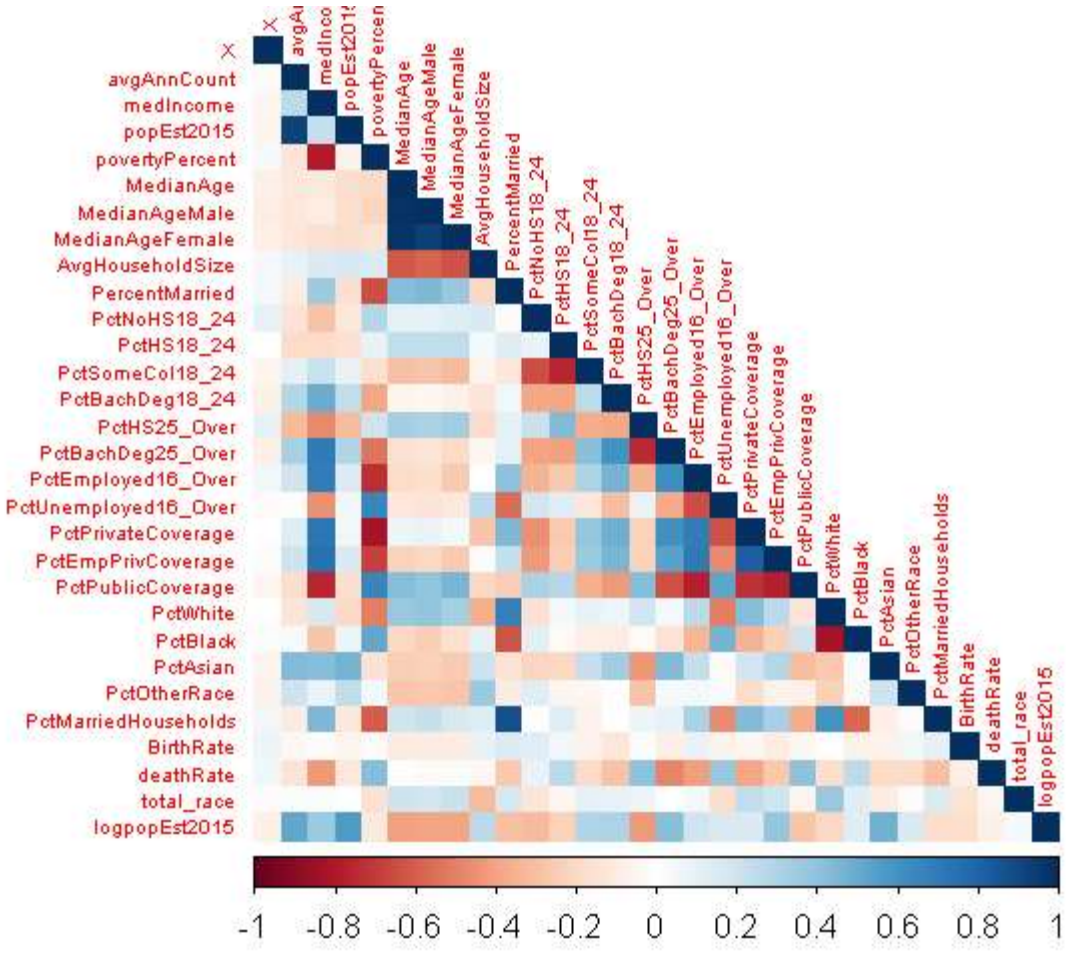
Given that our overall target is to better understand the death rate due to cancer, it probably makes most sense to look at this variable first. We see that it appears to follow more or less of a Gaussian curve *(note: refraining from implenting tests for normality until they are covered in course material).*

In [29]:
```
hist(data$deathRate, breaks = 40, main = "Death Rate")
```



To give guidance into what relationships in our data might be more interesting than others, we used the `corrplot` function from the `corrplot` package. This gives a visualization of the strength of the Pearson corrlation between various parameters. We note that we used a specific setting to only include values in which both observations were defined in calculating the Pearson coefficient (any `NA` values excluded). In this graphic, color __blue__ represents positive correlation factor while color __red__ represents negative correlation. White indicates there is no clear correlation between the variables.

In [30]:
```
# only include numerical columns in correlation matrix
numeric.cols <- as.logical(unlist(lapply(data, is.numeric)))
corr.matrix <- cor(data[,numeric.cols], use = "pairwise.complete.obs")

corrplot(corr.matrix, is.corr=T ,
         method = "color",
         type='lower',
         tl.cex = 0.5)
```



When looking at the corrplot, it is important to note that the colors only correspond to purely linear correlations in the data. If there are dynamics other than simple linear relationships, they may not have proper coloring to indicate that they are truly correlated. An example of this is to look at the difference in coloring of the `logpopEst2015` vs. `popEst2015`. The log transformed version of the data has stronger relationships with most variables, which indicates this might be a more useful variable to work with. In general, the `scatterplotMatrix` function from the `cor` package will give us better functionality to look at the distributions and relationships between a smaller amount of variables, but the above graphic allows for a better global view.

There are many "sanity checks" that we can gain from our data by looking at the corrplot. For instance, `medIncome` and `povertyPercent` are strongly negatively correlated. All of our parameters dealing with insurance make intuitive sense- the more public coverage, the less private coverage since people likely do not have 2 separate insurance plans. `PercentMarried` and `PctMarriedHouseholds` have a strong positive correlation, which seems logical. With sanity checks like these, we can ensure that parts of our data reflect reality.

Some interesting associations that are worth noting include the extremely high negative corrlation between `PctWhite` and `PctBlack`. While it makes sense these should be negatively correlated since they are different components of a whole, they show a much stronger relationship than any other two races. Additionally, the higher poverty rate areas and higher marriage percentage areas tend to have a higher percentage of black population and a lower percentage of whites. This shows some strong indication that segregation is prevalent in the counties in our dataset. In a different part of our data, education past a bachelor level is strongly correlated with both employment and private insurance coverage. Conversely, high poverty levels tend to be associated with higher public coverage, higher unemployment, and less bachelor-level education.

Our target variable, `deathRate`, appears to have some interesting correlations that should warrant further investigation. For positive correlations, `povertyPercent`, `PctHS25_Over`, `PctUnemployed16_Over`, and `PctPublicCoverage` appear to be important. For negative correlations, `medIncome`, `PctBachDeg25_Over`, and `PctEmployed16_Over` all seem important. Since ideally `PctEmployed16_Over` and `PctUnemployed16_Over` essentially tell the same information, we can look at only one of these. Even though they are less strong than the others mentioned, the insurance variables show that more public coverage tends to be associated with higher death rates, and more private coverage (whether through employers or self-bought) tends to be associated with lower death rates.

Subsetting of the high death rate counties

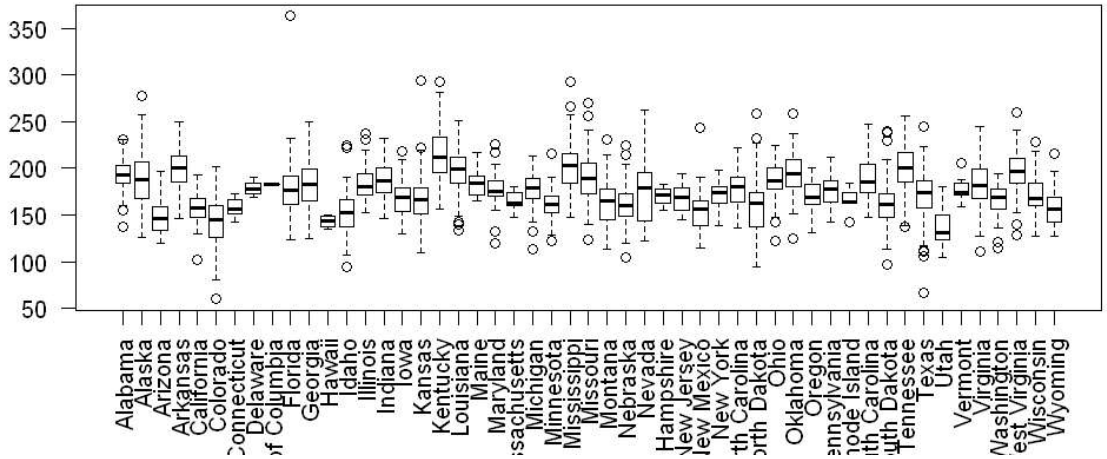
For further exploratory analysis, we want to define what constitutes a high death rate. We do this by using the upper bound inter-quartile range.

In [32]:
```
deathRatequrt3 <- quantile(data$deathRate, 0.75, na.rm=T)

data.higher <- subset(data, data$deathRate > deathRatequrt3)
print(paste0("There are ", nrow(data.higher), " observations in the highest qu
artile."))
print(paste0("The cutoff value is ", deathRatequrt3))

data.lower <- subset(data, data$deathRate < deathRatequrt3)
```

```
[1] "There are 761 observations in the highest quartile."
[1] "The cutoff value is 195.2"
```

We now look into the geography of the of the now-defined high death rate counties. From a boxplot, we see that the death rate varies state by state. A more detailed approach shows that the higher death rate states tend to be in the south, and Kentucky has the most high death rate counties by a wide margin.

In [63]:
```
options(repr.plot.width=8.5, repr.plot.height=4) # let this plot be wider
boxplot(deathRate~state, data = data, las = 2)
options(repr.plot.width=5, repr.plot.height=4) # set back to default settings
```



In [47]:
```
state.summ <- data.frame(summary(data$state))
names(state.summ) <- c("Counties_total")

# group with higher death rate
state.summ1 <- data.frame(summary(data.higher$state))
names(state.summ1) <- c("Counties_total")

state.summ$Counties_highDeathRate <- state.summ1$Counties_total

# Percentage of counties with higher death reate for each state
state.summ$Counties_HDR_Perc <- state.summ$Counties_highDeathRate/state.summ$C
ounties_total

head(state.summ[order(state.summ$Counties_HDR_Perc, decreasing = TRUE),])
```

	Counties_total	Counties_highDeathRate	Counties_HDR_Perc
Kentucky	120	91	0.7583333
Tennessee	95	57	0.6000000
Louisiana	64	38	0.5937500
Arkansas	75	44	0.5866667
Mississippi	82	47	0.5731707
West Virginia	55	28	0.5090909

The analysis above shows that higher death rates cross state lines, but tend to stay below the Mason-Dixon line. Kentucky seems to be in a group of its own, with roughly 76% of its counties in the high death rate range. The other high death rate states are bunched closer to the 50-60% range, so Kentucky stands out and warrants further analysis. A quick google search shows that Kentucky particularly has high rates of obesity and smoking along with a lack of screening. Further research into this could provide additional insight, but is outside of the scope of the data given for this report.

We now shift our focus to exploring if raw numbers of cancer incidents correspond to higher rates of cancer deaths. We use a state-wide aggregation approach to look at the ratio of cancer incidents to the total population, and see how these correspond to the death rate in the state.

In [59]:
```
sum_incdt <- data.frame(aggregate(data$avgAnnCount, by=list(Category=data$stat
e),FUN=sum))
names(sum_incdt) <- c("state","total_incident")
sum_incdt <- setDT(sum_incdt, keep.rownames = TRUE)[]
# tail(sum_incdt[order(sum_incdt$total_incident)],20)

#add the total population of state and the incident ratio
sum_pop <- data.frame(aggregate(data$popEst2015, by=list(Category=data$state),
FUN=sum))
sum_incdt$popEst <- sum_pop$x
sum_incdt$incdtRatio <- sum_incdt$total_incident/sum_incdt$popEst

# average state-wide death rate
sum_deathRate <- data.frame(aggregate(data$deathRate, by=list(Category=data$st
ate),mean))
sum_incdt$avgDeathRate <- sum_deathRate$x

head(sum_incdt[order(sum_incdt$incdtRatio, decreasing = TRUE),], 10)
```

rn	state	total_incident	popEst	incdtRatio	avgDeathRate
17	Kansas	200192.10	2831088	0.070712074	167.8343
24	Minnesota	170752.09	5489594	0.031104684	161.4816
29	Nevada	33365.35	2890845	0.011541729	177.4353
20	Maine	8214.00	1329328	0.006179062	183.1500
49	West Virginia	11130.00	1844128	0.006035373	196.7109
39	Pennsylvania	75909.00	12802503	0.005929231	175.4224
30	New Hampshire	7561.00	1330608	0.005682365	170.8800
46	Vermont	3548.00	626042	0.005667351	176.2714
7	Connecticut	20304.00	3590886	0.005654315	157.7125
18	Kentucky	24765.00	4425092	0.005596494	215.3158

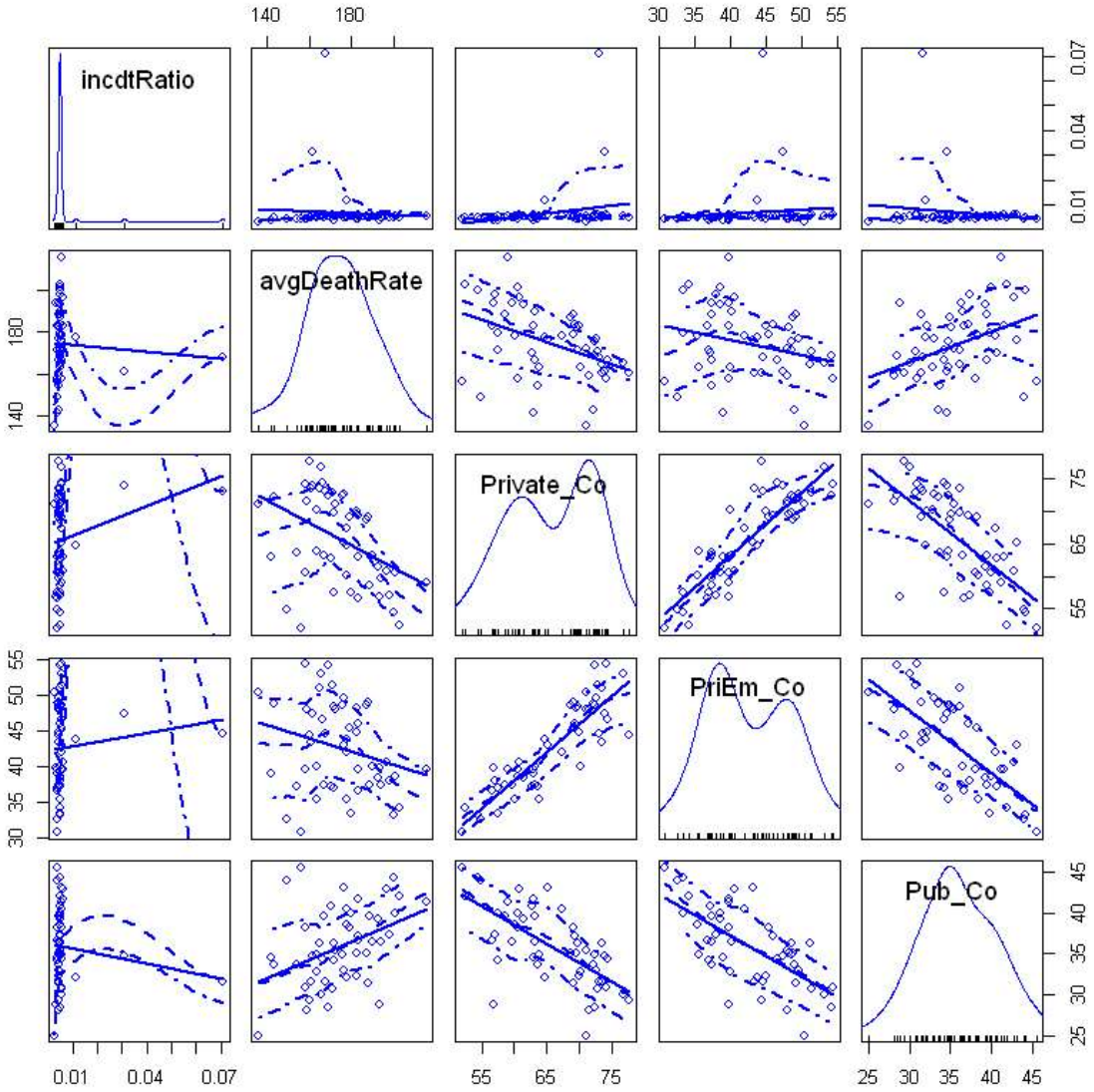
In [60]:
```
cor(sum_incdt[,c(5, 6)])
```

	incdtRatio	avgDeathRate
incdtRatio	1.00000000	-0.06621098
avgDeathRate	-0.06621098	1.00000000

Had cancer incidents been strongly related to cancer deaths, we would have expected to see Kentucky closer to the top of this list. We see that the correlation coefficient is close to zero, which shows that this relationship is not so strong. This was also seen in the corrplot earlier in the report. This suggests that the relationship between incidents and deaths is not the strongest factor at play and may not be the correct area to focus on for policy recommendation.

We now shift our focus to measures of treatment. In our dataset, this is best seen through the insurance parameters. In general, we would expect to see that higher percentages of private insurance plans would correspond to better treatment and therefore fewer cancer deaths.

In [67]:
```
# add to the table above with insurance coverage
# average private coverage by state
sum_PriCov <- data.frame(aggregate(data$PctPrivateCoverage, by=list(Category=d
ata$state),mean))

# average private employed coverage by state
sum_PriEmCov <- data.frame(aggregate(data$PctEmpPrivCoverage, by=list(Category
=data$state),mean))
# average public coverage by state
sum_PubCov <- data.frame(aggregate(data$PctPublicCoverage, by=list(Category=da
ta$state),mean))

sum_incdt$Private_Co <- sum_PriCov$x
sum_incdt$PriEm_Co <- sum_PriEmCov$x
sum_incdt$Pub_Co <- sum_PubCov$x

options(repr.plot.width=7, repr.plot.height=7)
scatterplotMatrix(~incdtRatio + avgDeathRate + Private_Co + PriEm_Co + Pub_Co,
 data = sum_incdt)
options(repr.plot.width=5, repr.plot.height=4)
```



We can see a few things by looking into the scatterplot matrix. First, the top 3 values of our incident ratio are outliers compared to the rest of the data. These correspond to Kansas, Minnesota, and Nevada. Next, there appears to be a bimodal distribution when looking at the private insurance coverage. We see a pronounced negative correlation between the death rate and all of the private insurance options, and a positive correlation with public insurance percentage. This gives some credence to the thought that better insurance may lead to better treatment for those with cancer, and ultimately lead to lower deaths.

In [75]:
```
sum_PovPerc <- data.frame(aggregate(data$povertyPercent, by=list(Category=data
$state),mean))

# average private employed coverage by state
sum_income <- data.frame(aggregate(data$medIncome, by=list(Category=data$state
),mean))
# average public coverage by state
sum_Employed <- data.frame(aggregate(data$PctEmployed16_Over, by=list(Category
=data$state),mean, na.rm = T))

sum_incdt$PovPerc <- sum_PovPerc$x
sum_incdt$income <- sum_income$x
sum_incdt$Employed <- sum_Employed$x

message("Poverty Percentage")
summary(sum_incdt$PovPerc)
message("median income")
summary(sum_incdt$income)
message("Percentage Employed")
summary(sum_incdt$Employed)
```

```
Poverty Percentage
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  9.838  12.685  15.081  15.875  18.625  25.399 
```

```
median income
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  35659   43438   50044   50575   55450   72135 
```

```
Percentage Employed
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  45.37   50.50   55.72   55.38   60.64   63.81 
```

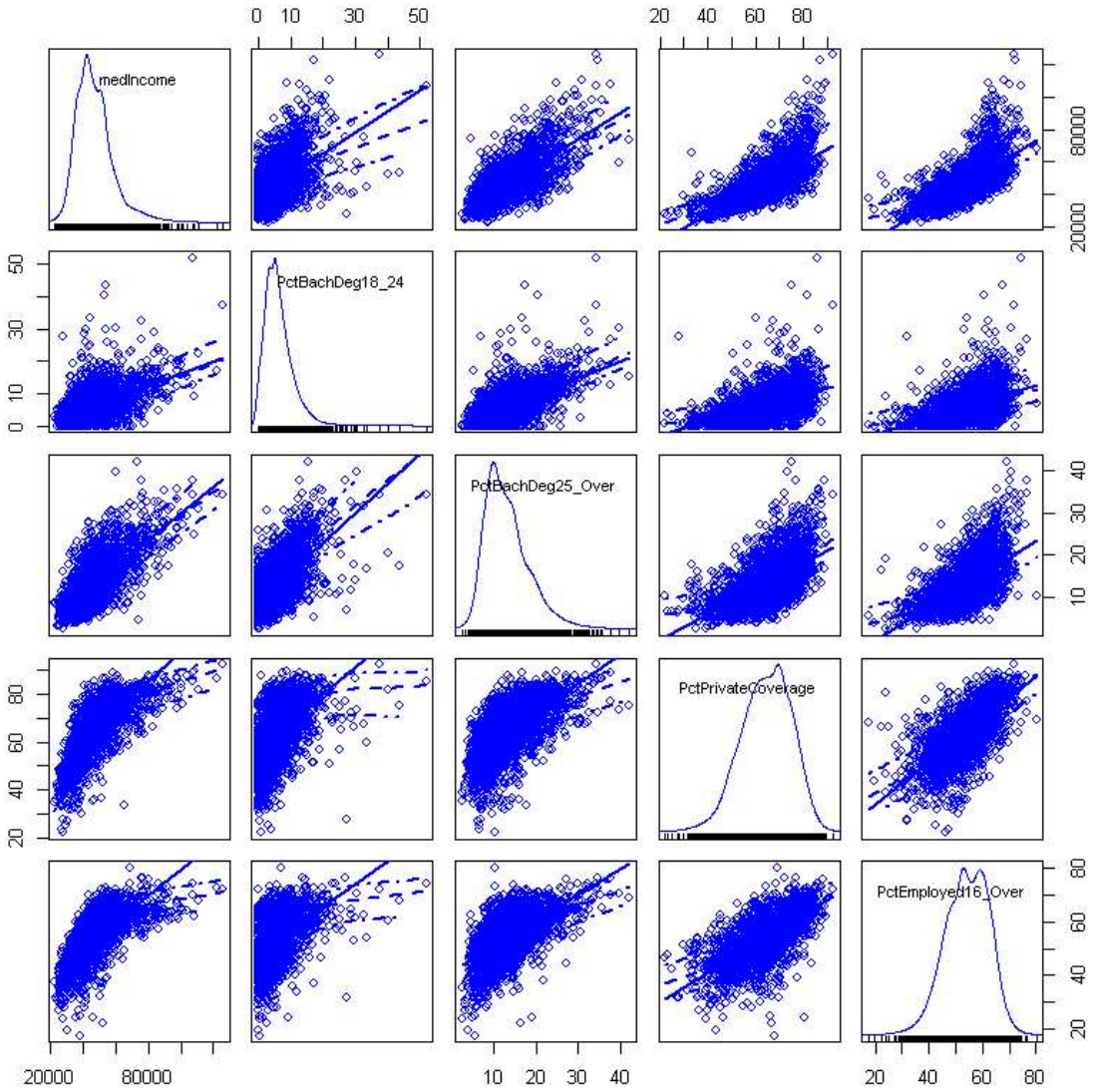
In [76]:
```
tail(sum_incdt[order(sum_incdt$avgDeathRate)],10)
```

rn	state	total_incident	popEst	incdtRatio	avgDeathRate	Private_Co	PriEm_C
26	Missouri	30939	6083672	0.005085580	189.6948	63.16348	40.07913
1	Alabama	24182	4750529	0.005090380	192.7286	61.48571	38.48730
2	Alaska	2512	706895	0.003553569	193.4167	56.79444	39.77222
37	Oklahoma	18555	3911338	0.004743901	193.9494	59.65325	37.29481
49	West Virginia	11130	1844128	0.006035373	196.7109	60.77636	40.61273
19	Louisiana	23242	4670724	0.004976102	197.6734	57.03906	37.90000
4	Arkansas	14762	2978204	0.004956679	200.0907	54.35333	33.29867
43	Tennessee	33284	6600299	0.005042802	200.8758	60.56737	38.53684
25	Mississippi	15007	2992333	0.005015150	202.8378	52.56707	34.20122
18	Kentucky	24765	4425092	0.005596494	215.3158	59.02000	39.64833

When aggregate by state to find the highest death rates, we find that all average incomes are below the mean income and higher than the median poverty percentage with the exception of Alaska. Additionally, all of the employment statistics for the high death rate states are below the mean value. Both of these facts support the hypothesis of more money means more availability to treatment.

As noted from the corrplot earlier in the analysis, it is important to note that private insurance and public insurance have strong relationships with employment and median income. This would seem to suggest that the ultimate thing effecting cancer death rate may be money, which is not independent from private insurance, employment, or education. We will take a deeper dive to see exactly the relationship with all of these variables.

In [72]:
```
options(repr.plot.width=7, repr.plot.height=7)
scatterplotMatrix(~medIncome + PctBachDeg18_24 + PctBachDeg25_Over + PctPrivat
eCoverage + PctEmployed16_Over, data = data)
options(repr.plot.width=5, repr.plot.height=4)
```



Across the board, we see clear positive correlations between all of these variables. We do not present any methods to prove causality in this analysis; therefore, we cannot say if any of these parameters directly lead to another.

Conclusions

Our analysis was intended to find relationships between demographic data and the death rate by cancer in the US. We hypothesized that there should be two different main aspects of the death rate: total overall cases of cancer, and quality of treatment.

Upon further analysis, there did not appear to be any sort of relationship between the cancer cases in a country (`avgAnnCount`) with a correlation coefficient of -0.066. We presume that treatment quality could be better with private coverage, and indeed there appears to be a tendency for cancer deaths to be lower in locations with higher private insurance coverages. These effects, however, could be due to a wide variety of parameters in our data. Income, employment and education levels have strong associations with private insurance coverage. Because we have not performed any statistical analysis to prove causality, we cannot prove that any one of these things causes the other or the cancer death rate.

For policy recommendations, one thing we have seen is that higher rates of public insurance leads to more cancer deaths. For whatever reason, there appear to be shortcomings in the treatment provided by public coverage. If we can improve the public coverage, this is probably the strongest independent lever we have to improve the death rate from cancer.