An Exploratory Analysis of Cancer Mortality and US Demographics by County

Tim Witthoefft, Mike Stackhouse, Craig Fujii, Wei Wang

1/29/2018

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 promote the understanding of 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 questions:

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 in 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 similarities 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

Variable Name	Variable Description	category
popEst2015	County population, estimated during 2015	Population/Birth
BirthRate*		Population/Birth
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

Variables with an * did not have a formal definition provided with the dataset. These definitions are inferred from the variable name.

Data Quality

The dataset was inspected using the following methods:

1. Calculating 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 variable. Note some scatterplots are not shown.

Missing Data

PctEmployed16_Over 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.

PctSomeCol18_24: 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 disproportionately large in comparison to the bulk of values within the variable. The largest of these values are as large as >38,000 with 10 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.

Univariate Analysis of Key Variables

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

Outcome deathRate and Average Annual Rate of Cancer

Since we do not know the units of cancer mortality rate (ie deathRate) variable, we have refrained from univariate analysis except to state that the data is normally distributed. In contrast, 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. A highly suspect bar is revealed when a histogram is performed on the transformed data. As discussed in the Data Quality section, we have elected to keep the data in the dataset.

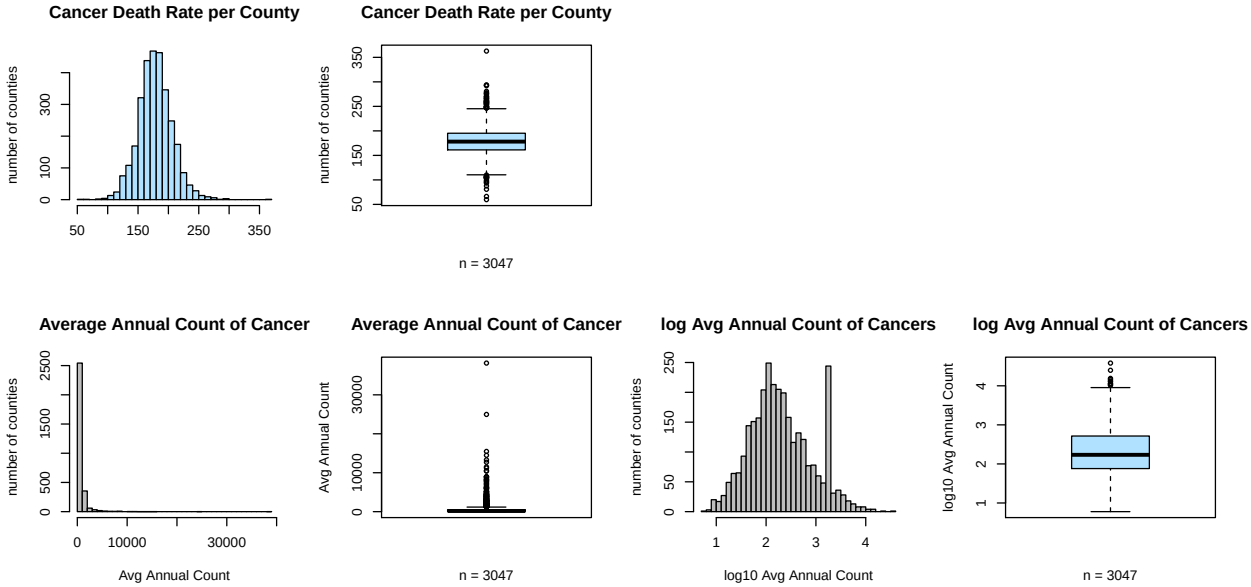
Summary Statistics:

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

```
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
```

```
layout(matrix(c(1, 2, 0, 0, 3, 4, 5, 6), 2, 4, byrow = T))
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, Birth Rate, and Geography

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 population helps visualize a distribution curve that looks normal. The birth rate is normally distributed with a very slight positive skew. We do not have the units for this variable so our analysis did not go further. We did perform a check to make sure there were no duplicate counties under the geography variable.

Summary Statistics:

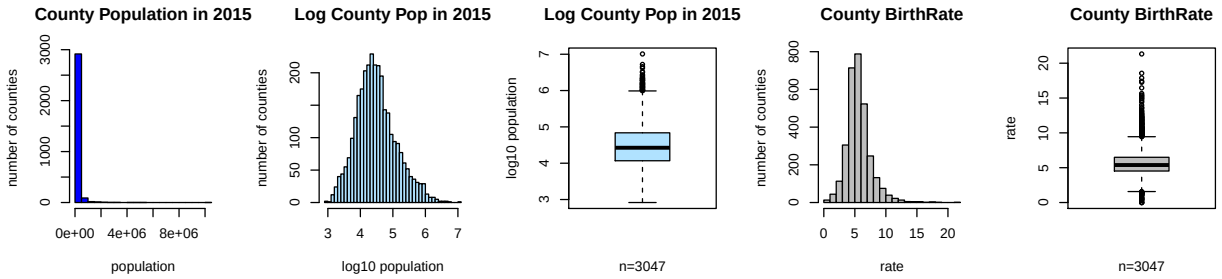
```
summary(data.frame(df$popEst2015, df$BirthRate))
```

```
##  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
```

```
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 = ""))
```




```
# 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)
```

Age

The 30 median age observations > 300 years were set to NA. The median age overall, for males, and for females per county all provide normal distributions. The median and mean age of females in the county is a 2-3 years greater than that for males - and both values are close to the mean as well. Barring the data issues found in the median age variable, the age data available did not raise any concerns.

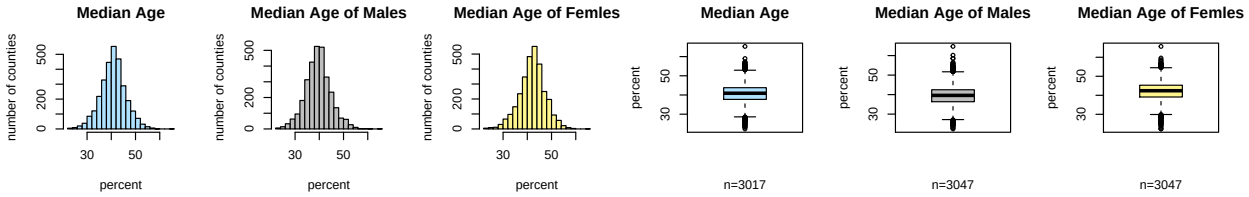
```
summary(data.frame(df$MedianAge, df$MedianAgeMale, df$MedianAgeFemale))
```

```
##   df.MedianAge    df.MedianAgeMale df.MedianAgeFemale
##  Min.   : 22.30   Min.   :22.40    Min.   :22.30
##  1st Qu.: 37.70   1st Qu.:36.35    1st Qu.:39.10
##  Median : 41.00   Median :39.60    Median :42.40
##  Mean   : 45.27   Mean   :39.57    Mean   :42.15
##  3rd Qu.: 44.00   3rd Qu.:42.50    3rd Qu.:45.30
##  Max.   :624.00   Max.   :64.70    Max.   :65.70
```

```
# sum((df$MedianAge>300), na.rm=TRUE) # identifies 30 observations median age
# greater than 300
df$MedianAge[df$MedianAge > 300] <- NA  #update the Median Age to NA
```

```
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

Most of the counties in the USA are predominantly white as shown by the left skewed histogram tallying the number of counties by percent white. As one would expect the minority ethnicities of Black, Asian, and Other Race are right skewed indicating some counties with sizeable minority populations where 96 counties had a Black population 50% or greater. With the data we have available, it appears that we'll be able to get the most insight into the impact of the prominence of Black and White races, while the small percentages of Asian and other races may not be large enough to understand the impact. The ethnicity data did not raise any significant concerns.

Summary Statistics:

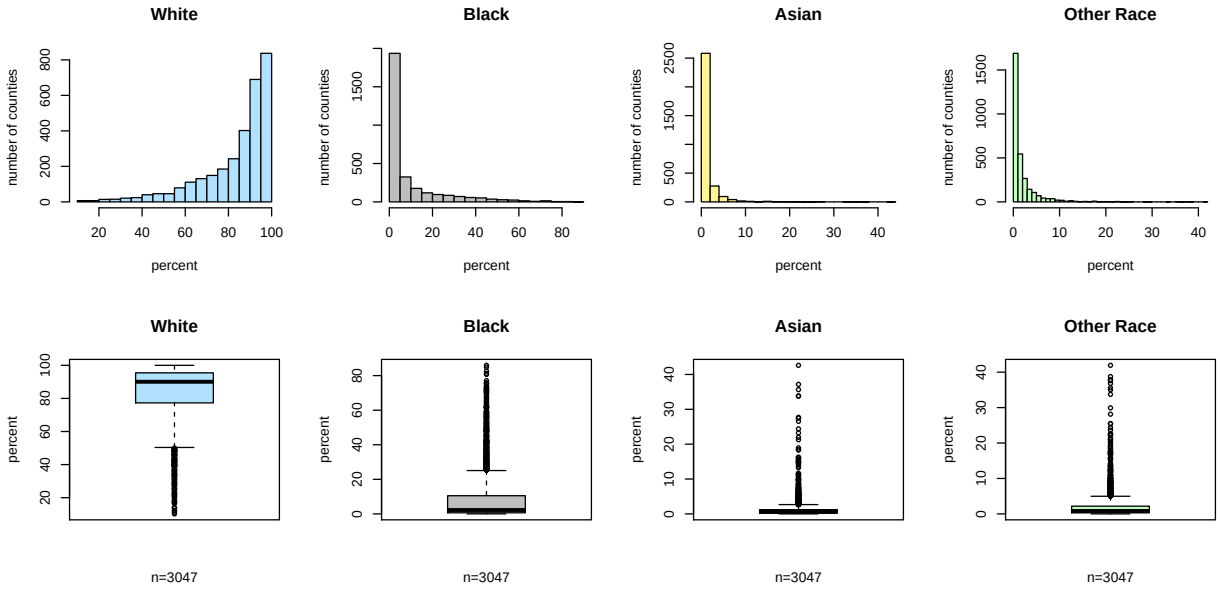
```
summary(data.frame(df$PctWhite, df$PctBlack, df$PctAsian, df$PctOtherRace), digits = 3)
```

```
##   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("White", "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.

Marital status appears to have a relatively normal distibution, with percent married having a slight left skew, though not a significant one. The summary statistics of these two variables are close together, so we should have fairly similar insights between both variables.

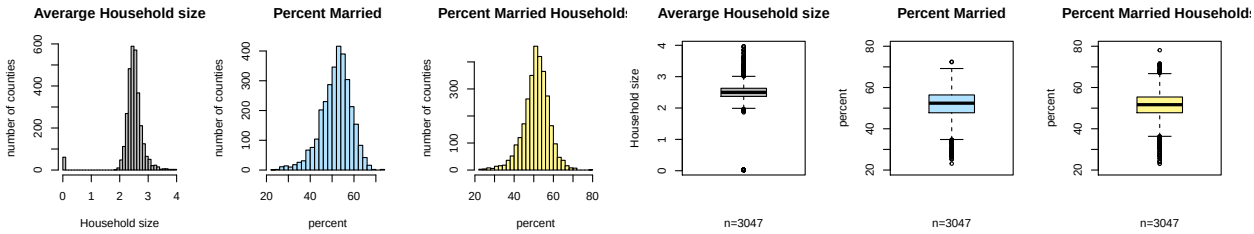
Summary Statistics:

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

```
##  df.AvgHouseholdSize df.PercentMarried df.PctMarriedHouseholds
##  Min.   :0.0221      Min.   :23.1      Min.   :23.0
```

```
##  1st Qu.:2.3700      1st Qu.:47.8      1st Qu.:47.8
##  Median :2.5000      Median :52.4      Median :51.7
##  Mean   :2.4797      Mean   :51.8      Mean   :51.2
##  3rd Qu.:2.6300      3rd Qu.:56.4      3rd Qu.:55.4
##  Max.   :3.9700      Max.   :72.5      Max.   :78.1
```

```
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$AvgHouseholdSize, main = title[2], breaks = 30, ylab = y[1], xlab = x[2],
    col = color[2])
hist(df$PercentMarried, main = title[1], breaks = 30, ylab = y[1], xlab = x[1], col = color[1])
hist(df$PctMarriedHouseholds, main = title[3], breaks = 30, ylab = y[1], xlab = x[1],
    col = color[3])
yl = c(20, 80)
boxplot(df$AvgHouseholdSize, main = title[2], col = color[2], ylab = y[3], xlab = paste(c("n=",
    sum(df$AvgHouseholdSize >= 0, na.rm = TRUE)), collapse = ""))
boxplot(df$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[3], col = color[3], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctMarriedHouseholds >= 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:

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

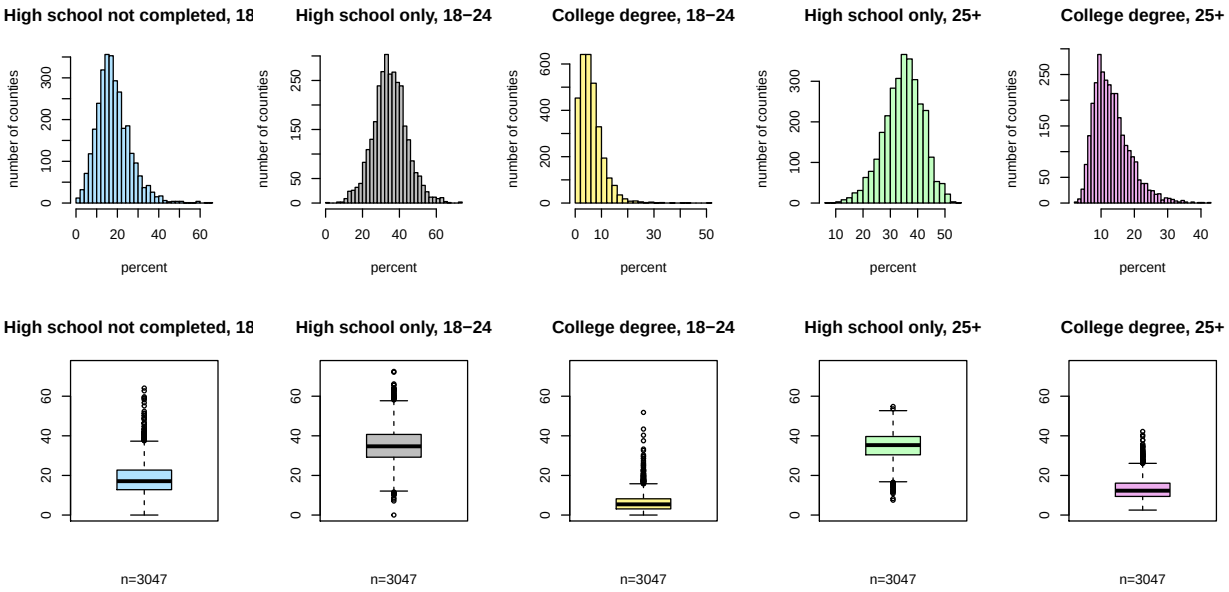
```
##  df.PctNoHS18_24 df.PctHS18_24  df.PctHS25_Over df.PctBachDeg18_24
##  Min.   : 0.00   Min.   : 0.0   Min.   : 7.50   Min.   : 0.000
##  1st Qu.:12.80   1st Qu.:29.2   1st Qu.:30.40   1st Qu.: 3.100
##  Median :17.10   Median :34.7   Median :35.30   Median : 5.400
##  Mean   :18.22   Mean   :35.0   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 (34%). 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.

```
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 right 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 left skew, with highly percentages much more common than low percentages.

The variable *PctBachDeg18_24* was the most extreme distribution with a far-right skew. This is likely due to the definition 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

There are three variables that are categorized under Insurance Coverage: PctPrivateCoverage, PctEmpPrivCoverage and PctPublicCoverage.

Summary Statistics:

```
summary(data.frame(df$PctPrivateCoverage, df$PctEmpPrivCoverage, df$PctPublicCoverage),
    digits = 4)
```

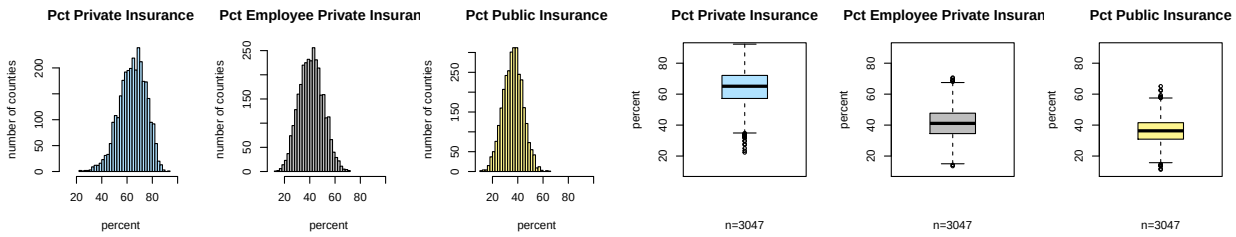
```
##  df.PctPrivateCoverage df.PctEmpPrivCoverage df.PctPublicCoverage
##  Min.   :22.30         Min.   :13.5          Min.   :11.20
##  1st Qu.:57.20         1st Qu.:34.5          1st Qu.:30.90
##  Median :65.10         Median :41.1          Median :36.30
##  Mean   :64.35         Mean   :41.2          Mean   :36.25
##  3rd Qu.:72.10         3rd Qu.:47.7          3rd Qu.:41.55
##  Max.   :92.30         Max.   :70.7          Max.   :65.10
```

The summary above shows that none of the three variables have missing values. The percentage of people who have private coverage is the largest among the three.

```
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)
xl = c(10, 100)
hist(df$PctPrivateCoverage, main = title[1], breaks = 30, ylab = y[1], xlab = x[1],
    xlim = xl, col = color[1])
hist(df$PctEmpPrivCoverage, main = title[2], breaks = 30, ylab = y[1], xlab = x[1],
    xlim = xl, col = color[2])
hist(df$PctPublicCoverage, main = title[3], breaks = 30, ylab = y[1], xlab = x[1],
    xlim = xl, 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 = ""))
```



The histograms and the boxplots show that the PctPrivateCoverage variable is slightly skewed left. However it's close to the to normal distribution and the other two variables has normal distribution as well. So no transformation needs to be done and all three variables can be included in the analysis.

Employment

Other than the 152 missing data observations, the two variables describing employment of residents 16 years or older show a normal distribution. Percent residents employed in a county has a mean and median near 54 with a first quartile at 48.6% indicating 75% of the counties had close to half their residents employed. In comparison to unemployment percentages, 75% of the counties had an unemployment rate below 9.7%. While not the focus of this analysis, it is noted that the two rates per county do not add up to 100% which may imply an additional category such as retired (i.e. not employed and not seeking employment). The employment data did not raise any concerns.

Summary Statistics:

```
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
```

```
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)
xl = c(0, 80)
hist(df$PctEmployed16_Over, main = title[1], breaks = 30, ylab = y[1], xlab = x[1],
    xlim = xl, col = color[1])
hist(df$PctUnemployed16_Over, main = title[2], breaks = 30, ylab = y[1], xlab = x[1],
    xlim = xl, 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$PctUnemployed16_Over, main = title[2], col = color[2], ylab = y[2], ylim = yl,
    xlab = paste(c("n=", sum(df$PctHS18_24 >= 0, na.rm = TRUE)), collapse = ""))
```

Income and Poverty level

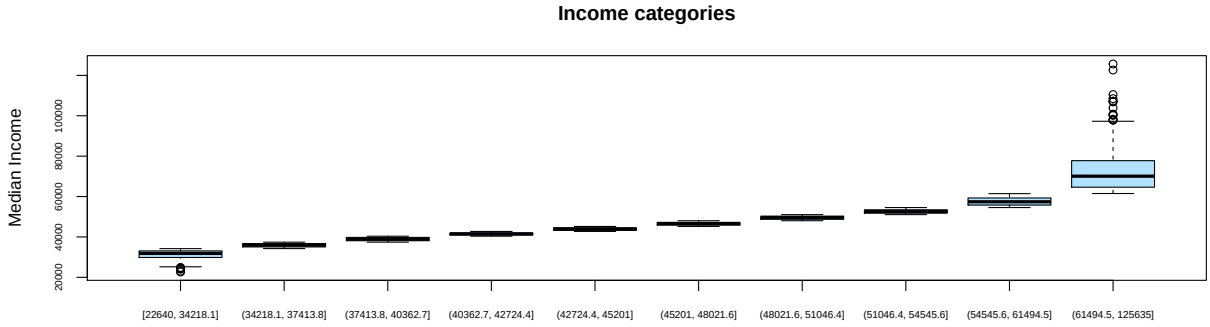
Median income was provided for each county and assigned a bin containing 302-306 observations. We confirmed that the 10 income bins are based on the deciles of the median income of the dataset. Median income has a right skew indicating there are counties much greater median income above the mean.

Percent of residents below the poverty level was also included the dataset and presents a right skew as well indicating select counties have sizeable portion of their residents in this condition. The poverty level is something all counties must address where if one interprets the inter-quartile range (IQR) of the boxplot, 50% of the counties have 12.2- 20.4% of their residents considered impoverished. This means that the 25% most impoverished counties have 20.4% to as high as 47.4% of their residents in poverty. In terms of analysis, the income and poverty data did not raise any concerns.

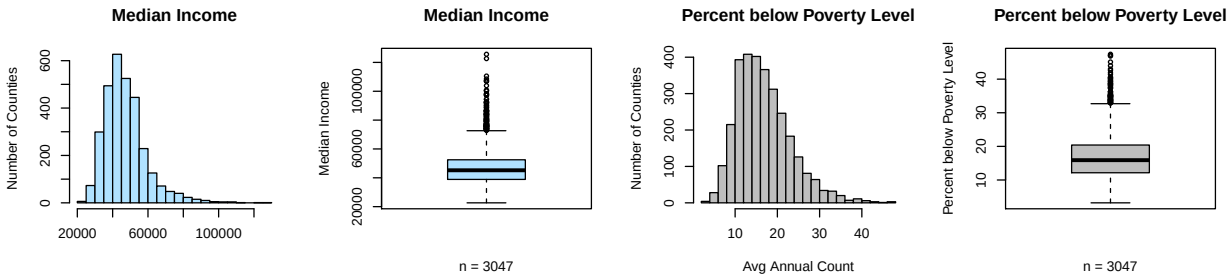
```
df$binnedInc <- factor(df$binnedInc, levels = c("[22640, 34218.1]", "(34218.1, 37413.8]",
    "(37413.8, 40362.7]", "(40362.7, 42724.4]", "(42724.4, 45201]", "(45201, 48021.6]",
    "(48021.6, 51046.4]", "(51046.4, 54545.6]", "(54545.6, 61494.5]", "(61494.5, 125635]"))
summary(df$binnedInc)
```

```
##   [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
```

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




```
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

All independent varialbes were analyze for linear correleation with the deathRate. Those with a correlation were investigated further using a scatterplot or a boxplot comparison.

Table below show below shows the correlation coefficients for r

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

vars <- c()
cors <- c()
cats <- c()
```

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

        if (i %in% c("MedianAge", "MedianAgeMale", "MedianAgeFemale")) {
            cats <- c(cats, "Age")
        } else if (i %in% c("PctWhite", "PctBlack", "PctAsian", "PctOtherRace")) {
            cats <- c(cats, "Ethnicity")
        } else if (i %in% c("AvgHouseholdSize", "PercentMarried", "PctMarriedHouseholds")) {
            cats <- c(cats, "Household/Marital")
        } else if (i %in% c("PctNoHS18_24", "PctHS18_24", "PctSomeCol18_24", "PctBachDeg18_24",
            "PctHS25_Over", "PctBachDeg25_Over")) {
            cats <- c(cats, "Education")
        } else if (i %in% c("PctPrivateCoverage", "PctEmpPrivCoverage", "PctPublicCoverage")) {
            cats <- c(cats, "Insurance")
        } else if (i %in% c("PctEmployed16_Over", "PctUnemployed16_Over")) {
            cats <- c(cats, "Employment")
        } else if (i %in% c("medIncome", "povertyPercent", "binnedInc")) {
            cats <- c(cats, "Income")
        } else if (i %in% c("avgAnnCount")) {
            cats <- c(cats, "Cancer Incidence")
        } else if (i %in% c("popEst2015", "BirthRate")) {
            cats <- c(cats, "Population/Birth")
        } else {
            cats <- c(cats, "????")
        }
    }
}

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

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

Table 2: Correlation Coefficients of Independent Variables to Death Rate

Variable	Category	Correlation Coefficient
MedianAge	Age	-0.0043
MedianAgeFemale	Age	0.0120
MedianAgeMale	Age	-0.0219
avgAnnCount	Cancer Incidence	-0.1435
PctNoHS18_24	Education	0.0885
PctSomeCol18_24	Education	-0.1887
PctHS18_24	Education	0.2620
PctBachDeg18_24	Education	-0.2878
PctHS25_Over	Education	0.4046
PctBachDeg25_Over	Education	-0.4855
PctUnemployed16_Over	Employment	0.3784
PctEmployed16_Over	Employment	-0.4120
PctWhite	Ethnicity	-0.1774

Variable	Category	Correlation Coefficient
PctAsian	Ethnicity	-0.1863
PctOtherRace	Ethnicity	-0.1899
PctBlack	Ethnicity	0.2570
AvgHouseholdSize	Household/Marital	-0.0369
PercentMarried	Household/Marital	-0.2668
PctMarriedHouseholds	Household/Marital	-0.2933
povertyPercent	Income	0.4294
PctEmpPrivCoverage	Insurance	-0.2674
PctPrivateCoverage	Insurance	-0.3861
PctPublicCoverage	Insurance	0.4046
BirthRate	Population/Birth	-0.0874

Marriage and Household vs. Death Rate

Marriage seems to have a slight negative relationship with cancer death rates. The dataset had two variables to explore: *PctMarried* and *PercentMarriedHouseholds.* Unsurprisingly, these variables have correlations very close to each other, with *PctMarried* having -0.293 and *PctMarriedHouseholds* having -0.293. This may suggest that partner support leads to better outcomes with cancer mortality and would be worth investigating further. Average household size did not have any correlation with death rate.

```
par(mfrow = c(1, 2))
r_cor = round(cor(df$deathRate, df$PctMarriedHouseholds, use = "complete.obs"), 5)
scatterplot(df$PctMarriedHouseholds, df$deathRate, xlab = "Percentage of Households Married",
    ylab = "Death Rate", main = "Cancer death rate vs. Percent Households Married",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))

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

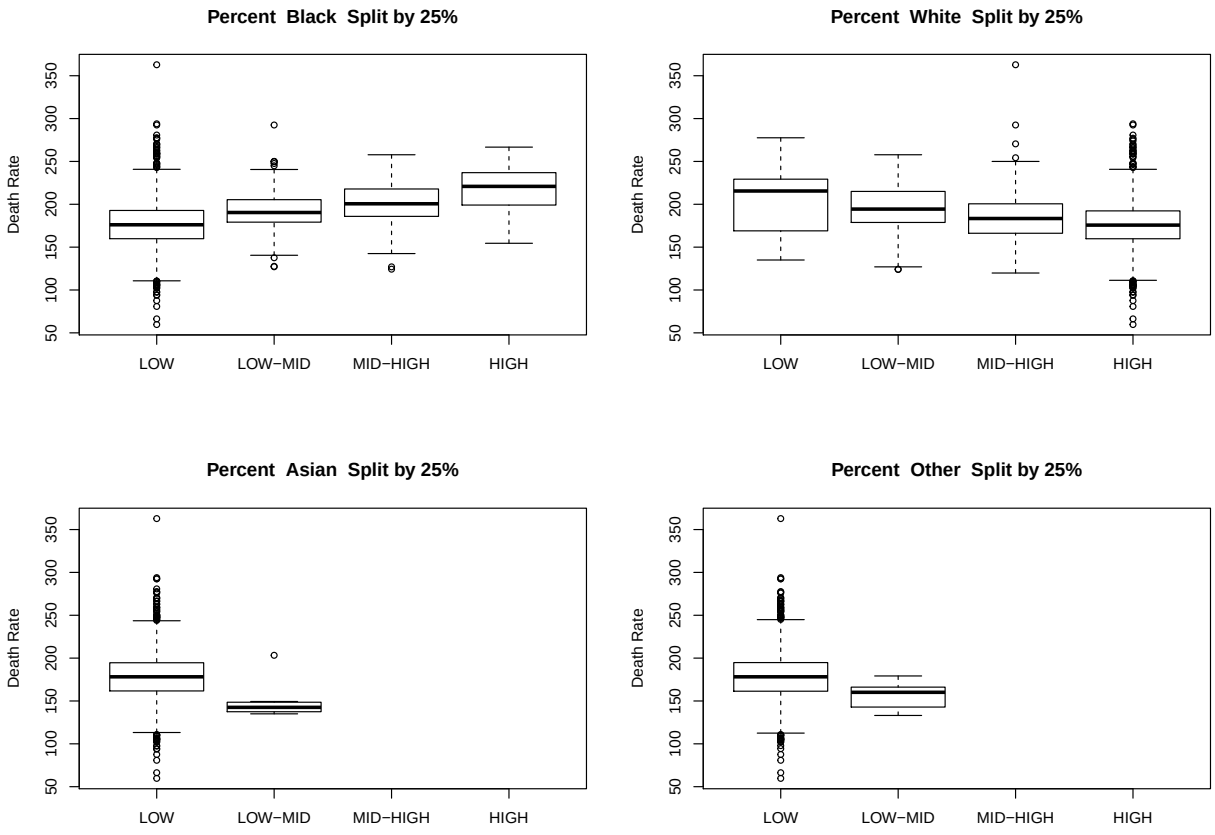
Race vs. Death Rate

Race has some interesting interactions with Death Rate in the fact that relationships differ with the prominence of a single race in a county. All races except for Black have a negative correlation, though a weak one, with the death rate. Conversely, the Black has a positive correlation that is stronger than the other correlations.

We can gain the most insight looking at white and black communities, while there is not enough data to gain any valuable insight into prominent Asian communities or communities of other races.

The race percentage variables have significant skew to their distributions, so to better represent this data, have chosen to display them in box plots with binned percentages of each quartile. This allows us to examine how the mean changes through each quartile and better see the relationship as the concentration of each race in a county grows. Looking at these results, it appears that the death rate grows as the black percentage grows, while it decreases slightly as the white percentage grows. This is one area that has the potential to look at secondary effects.

```
par(mfrow = c(2, 2))
bplot_race <- function(race, racec) {
    race_bin = cut(race, breaks = c(0, 25, 50, 75, 100), label = c("LOW", "LOW-MID",
        "MID-HIGH", "HIGH"))
    boxplot(df$deathRate ~ race_bin, data = df, main = paste("Percent ", racec, " Split by 25%"),
        ylab = "Death Rate")
}
bplot_race(df$PctBlack, "Black")
bplot_race(df$PctWhite, "White")
bplot_race(df$PctAsian, "Asian")
bplot_race(df$PctOtherRace, "Other")
```



Age vs. Death Rate

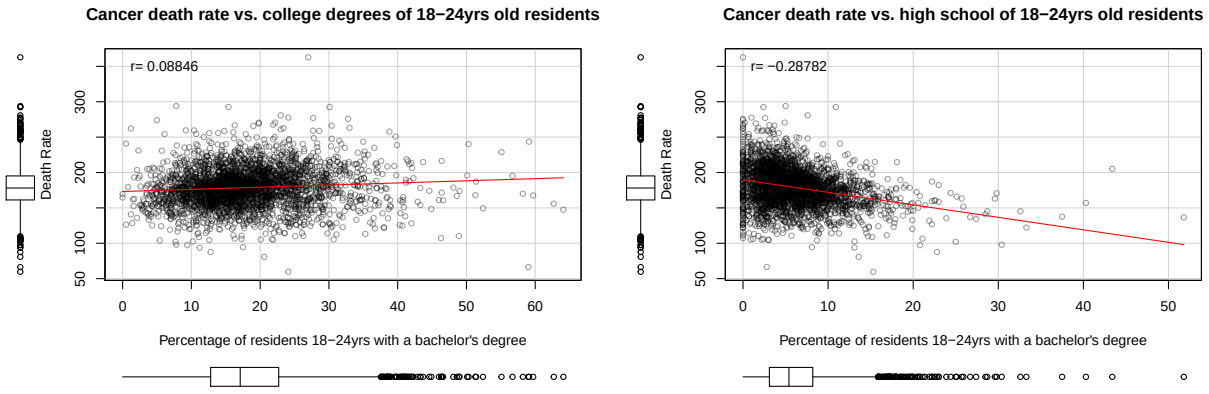
The interesting result about age compared to death rate is that the correlation is rather weak. One might guess that as a community ages, the death rate would grow. Our sample contradicts this assumption, with the Median Age having a correlation of only -0.00429. We see similar values in Median Age of Males (-0.02193) and Median Age of Females (0.01205).

Education

The first thing to note about education 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.

```
par(mfrow = c(1, 2))
r_cor = round(cor(df$deathRate, df$PctNoHS18_24, use = "complete.obs"), 5)
scatterplot(df$PctNoHS18_24, df$deathRate, xlab = "Percentage of residents 18-24yrs with a bachelor's d
    ylab = "Death Rate", main = "Cancer death rate vs. college degrees of 18-24yrs old residents",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))

r_cor = round(cor(df$deathRate, df$PctBachDeg18_24, use = "complete.obs"), 5)
scatterplot(df$PctBachDeg18_24, df$deathRate, xlab = "Percentage of residents 18-24yrs with a bachelor'
    ylab = "Death Rate", main = "Cancer death rate vs. high school of 18-24yrs old residents",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))
```

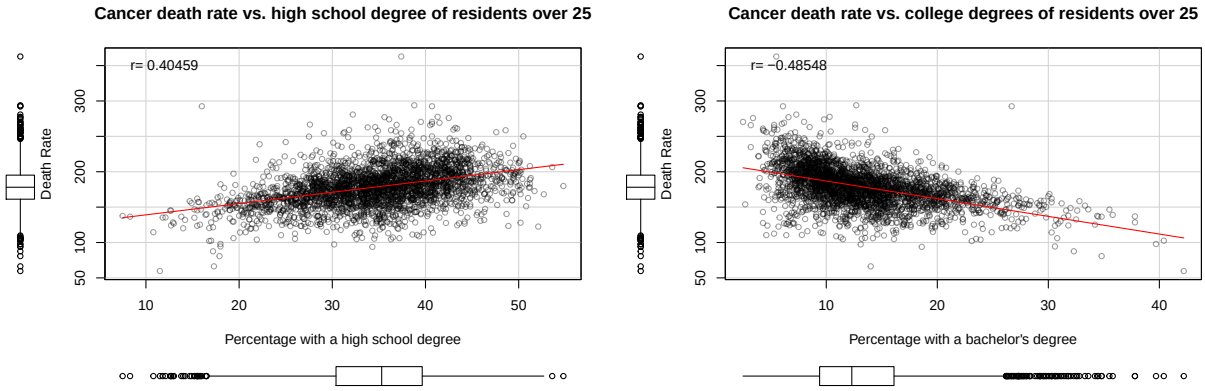


Observing only the two "over 25 age" variables, we see two distinctly different trends. The first - percentage of residents with a college degree - shows a moderately decreasing trend with a negative correlation of -0.4854773. The second variable, *PctHS25_Over*, conversely shows an increasing trend with a correlation of 0.4045891.

```
par(mfrow = c(1, 2))
r_cor = round(cor(df$deathRate, df$PctHS25_Over, use = "complete.obs"), 5)
scatterplot(df$PctHS25_Over, df$deathRate, xlab = "Percentage with a high school degree",
    ylab = "Death Rate", main = "Cancer death rate vs. high school degree of residents over 25",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))

r_cor = round(cor(df$deathRate, df$PctBachDeg25_Over, use = "complete.obs"), 5)
scatterplot(df$PctBachDeg25_Over, df$deathRate, xlab = "Percentage with a bachelor's degree",
```

```
    ylab = "Death Rate", main = "Cancer death rate vs. college degrees of residents over 25",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))
```



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.

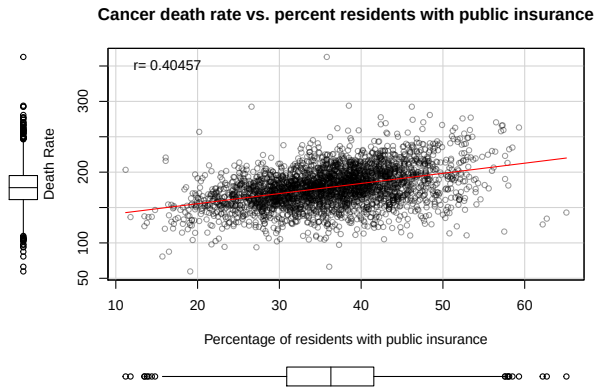
Insurance

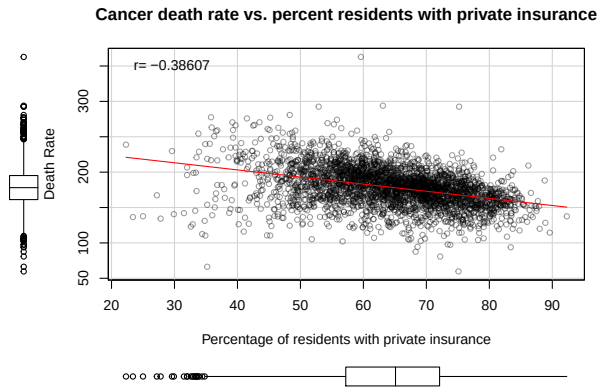
The correlation chart below indicates that the Percentage Public Coverage variable has the strongest correlation with the Death Rate whereas the relationship between the Percentage Private Coverage variable and the Death Rate agrees with the relationship between the Percentage Employer Private Coverage and the Death Rate.

Death rate Scatterplot

```
par(mfrow = c(1, 2))
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. percent residents with public insurance",
    legend("topleft", bty = "n", legend = paste("r=", r_cor)), col = c("red", "green",
        rgb(0, 0, 0, 100, maxColorValue = 255)))

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



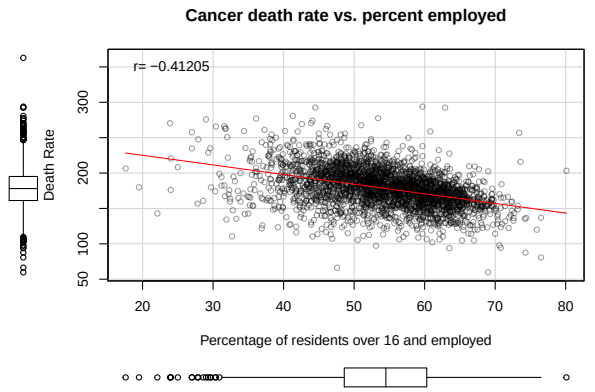


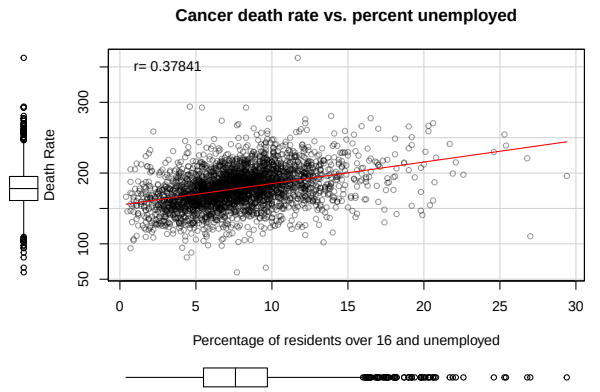
Employment

Employments has appears to have one of the stronger relationships with the death rate. This would make sense, as employment offers better access to health care (and also is another area to look for secondary effects). The results here are fairly clear and follow a logical trend; the higher the percent employed, the lower the death rate. The higher the percent unemployed, the higher the death rate.

```
par(mfrow = c(1, 2))
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. percent employed", legend("topleft",
        bty = "n", legend = paste("r=", r_cor)), col = c("red", "green", rgb(0, 0,
        0, 100, maxColorValue = 255)))

r_cor = round(cor(df$deathRate, df$PctUnemployed16_Over, use = "complete.obs"), 5)
scatterplot(df$PctUnemployed16_Over, df$deathRate, xlab = "Percentage of residents over 16 and unemploy
    ylab = "Death Rate", main = "Cancer death rate vs. percent unemployed", legend("topleft",
        bty = "n", legend = paste("r=", r_cor)), col = c("red", "green", rgb(0, 0,
        0, 100, maxColorValue = 255)))
```





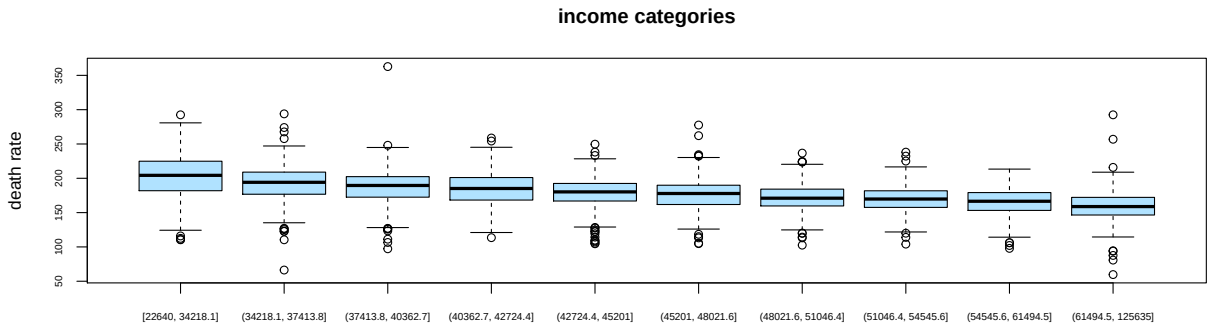
Income

The binned income shows a downward trend in death rate as the binned median incomes increases. This has a similar relationship that one might expect with employment. Higher income could imply better access to

healthcare and health related resources, so this opens another area that could be examined for secondary effects.

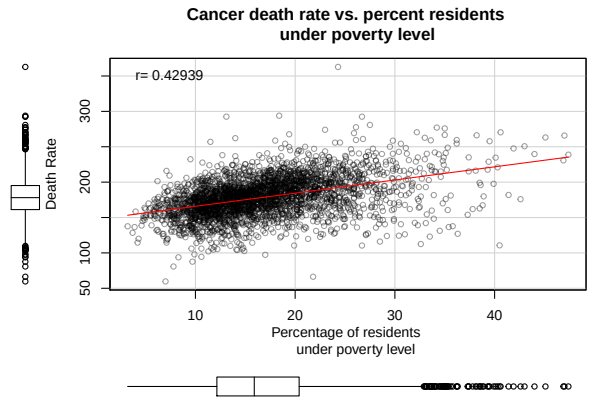
Median income

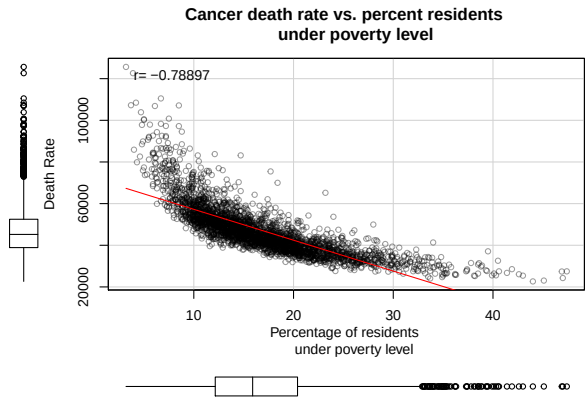
```
boxplot(df$deathRate ~ df$binnedInc, data = df, cex.axis = 0.58, main = "income categories",
    ylab = "death rate", col = "lightskyblue1")
```




```
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_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 correlation and Secondary Effects

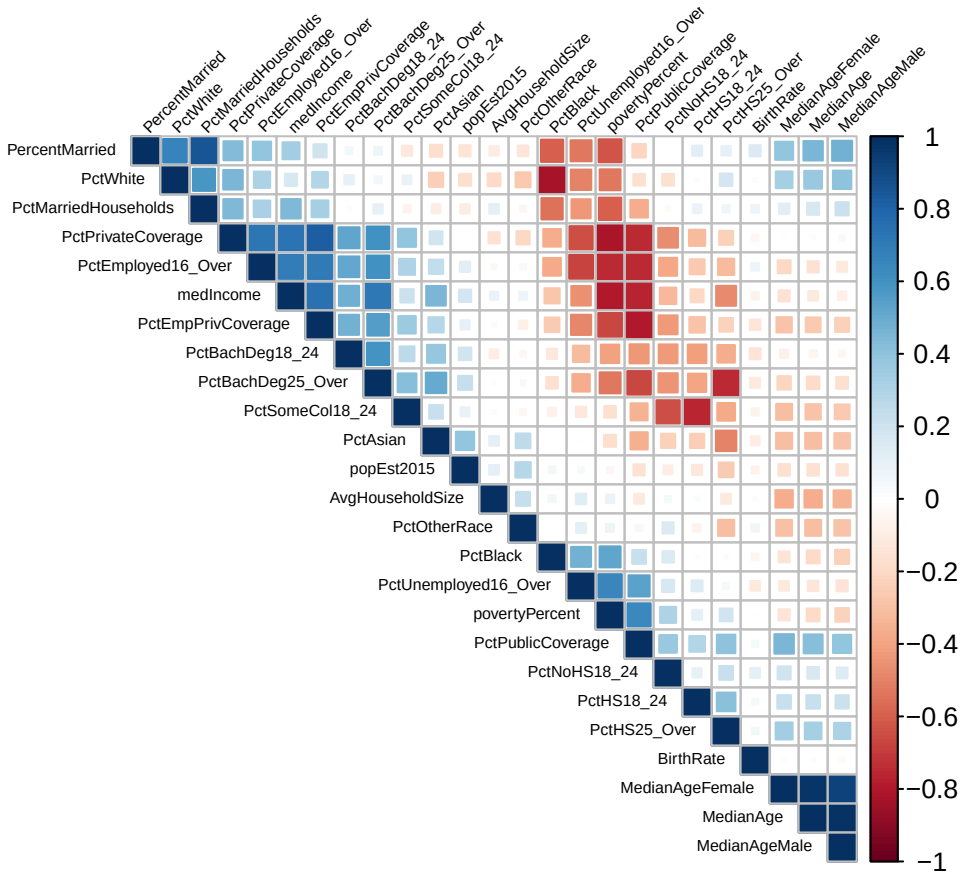
This is a challenging dataset due to presence of multiple potentially confounding variables - variables that can influence both the dependent (target) variable, as well as other independent variables.

To help identify these types of variables, the following correlation matrix was created for variables in the dataset. This matrix excludes *deathRate* with the intention of only showing correlation between the chosen independent variables. The matrix was intentionally ordered to group similarly-correlated variables together.

Independent Variable Correlation Matrix

```
cr <- cor(df[, c("popEst2015", "BirthRate", "MedianAge", "MedianAgeMale", "MedianAgeFemale",
    "PctWhite", "PctBlack", "PctAsian", "PctOtherRace", "AvgHouseholdSize", "PercentMarried",
    "PctMarriedHouseholds", "PctNoHS18_24", "PctHS18_24", "PctHS25_Over", "PctSomeCol18_24",
    "PctBachDeg18_24", "PctBachDeg25_Over", "PctPrivateCoverage", "PctEmpPrivCoverage",
    "PctPublicCoverage", "PctEmployed16_Over", "PctUnemployed16_Over", "medIncome",
    "povertyPercent")], use = "complete.obs")

corrplot(corr = cr, type = "upper", tl.col = "black", tl.cex = 0.5, tl.srt = 45,
    method = "square", order = "AOE")
```



The correlation matrix shows in multiple ways that each variable of the dataset cannot be considered a truly *independent* variable. For example, the variables concerning insurance and education all correlate with the other variables of that subgroup. For example, the percent of county residents on either private or public insurance are highly correlated - a strong percentage in one shows a smaller percentage in the other. Likewise, high percentages of high school only education result in lower percentages of college educated residents, and vice versa. This observation suggests that no one variable is necessarily more important that the other within the subgroup, but rather these variables are measuring multiple possible outcomes on a range of related

results (e.g. no high school education, high school only, college educated, etc.)

The variables that consistently showed small correlations with other variable included BirthRate, popEst2015 (population), and AvgHouseholdSize.

The analysis of key relationships identified six areas where a correlation to the target variable deathRate can be observed. Below are further discussions on each grouping in preparation of identifying characteristics associated with cancer death rate to inform social intervention.

Education

To further elaborate, the two different age groups 18-24 and 25 and older did yield different correlations and as mentioned one may be a secondary effect of the other. While the 25 or older percentages may have yielded higher correlations, these variables may be secondary to the 18-24 percentages. For example, if a county school district has a tradition of good schools preparing their students for college, then the high percentages in the PctBachDeg18_24 variable may be more informative to identify interventions that can be applied to the counties with low percentages in the PctBachDeg18_24.

Poverty and Income

Percent poverty and median income appear to be inversely related to each other in that the greater one's income the less likely one is in poverty. Poverty is the variable positively correlated with death rate and may be a good candidate variable to identify interventions to decrease this correlation. Variables positively correlated to Poverty include percent unemployed and percent Black. Both may be secondary variables to poverty because if one does not have a job one is likely to be at greater risk of becoming impoverished and being black may be related to discrimination and less opportunities. Median Income is has several variables positively correlated with it and includes percent employed, percent private coverage, percent married household, percent college degree, and percent Asia. Having more positively correlated variables may lead to more secondary effects to investigate and may suggest poverty may be the better choice variable to inform decisions.

Insurance

Public insurance is negatively correlated with private insurance, median income, and percent employed and percent with college degrees. This might be related to the converse of public insurance being private insurance where those with private insurance is secondary to being able to afford private insurance, having good paying job which may have been enabled with a college degree. In contrast, public insurance is associated with poverty, birth rate, and age. Poverty may be one underlying reason because impoverished individuals may not be able to afford quality healthcare. The positive correlation with age may indicate medicare if that is considered a public coverage. However, age in itself had not correlation with the death rate which may point to that the poverty association may mean that those who are dying of cancer have public coverage and are poor. Birth rate suggests the more kids one has the more likely one is to be on public insurance.

Race

PctBlack and PctWhite have opposite trending correlations and are negatively correlated to each other meaning that they are likely dependent on each other. In other words, a high percent of white population will have a low percent of black and vice versa. Another way to understand this is a high absence of a white population is associated with a high percent of a black population. In this manner only one variable, PctBlack may be sufficient enough, for analysis.

Percent Married Household

Percent Married Household is correlated with percent married, percent white, percent employed, percent employed, median income, percent employee private coverage. These underlying variables may enable the indvidual to enter marriage and hold together a household which would suggest that percent married house hold is a secondary effect variable to these variables.

Summary and Conclusion

The dataset contained 3047 US counties and our team found this to have high enough quality to provide insights to answer our research questions despite some the missing data and data points suspected to be erroneous. Below is summary of the issues

2285 missing values for variable PctSomeCol18_24 (75% of observations) 30 erroneous MedianAge values (1%) 61 potentially erroneous AvgHouseholdSize (2%) 206 potentially erroneous avgAnnCount (7%) 152 missing values PctEmployed16_Over (5%)

The 75% missing values in PctSomeCol18_24 did weaken this variable's value to in our analysis. MedianAge and AvgHouseholdSize did not show any associations with deathRate. The avgAnnCount was not focused on in this study because death rate was the target variable and not incidence of cancer. The PctEmployed16_Over did prove to be a variable showing association with deathRate despite the missing values.

Per our first research question key county level characteristics associated with mortality rates from cancer, our team found 6 areas in the dataset.

Education: Percent of residents with a HS degree positively correlated with greater cancer death rates. In contrast, higher percentages of residents with a bachelor's degree correlated with lower death rates. This was observed in the 18-24 residents (r, BS: -0.2878 vs HS: 0.2620) and more pronounced in the 25 years or older residents (r, BS -0.4855 vs HS: 0.4046).

Poverty Percent residents in poverty had a moderate correlation with death rate (r = 0.4294). Not surprising, median income was negatively correlated (r =-0.4286) Insurance - Percent of residents in a county with public insurance positively correlated with cancer deathrate (r = 0.4046) in contrast to those counties with higher percent of residents with private insurance (r= -0.3861).

Insurance Employee private coverage had a less pronounced negative correlation. We identified the percent of residents with employee private coverage to be a possible secondary effort included in the percent with private coverage.

*Employment*Percent employment was had a moderate negative correlation (r= -0.4120) and not surprisingly percent unemployment was positively correlated (r= 0.3784). Percent employment may be a secondary effect of private insurance because employed individual may have access to employee benefits including health insurance. We noted employment as a secondary effect because percent residents with private coverage had a greater correlation than percent with employee private insurance.

Race showed a weak positive correlation for counties with percent black with death rate (r = 0.2570) and a negative correlation for the percent white in a county. <<The weak correlation indicates it is likely a secondary effect as percent black was associated with poverty and public insurance.

*Marriage*Percent Married showed a weak negative correlation (r=0.2933) suggesting that being in a relationship may be helpful in preventing cancer death.

County Characteristics to inform action

Our findings did enable us to identify county level characteristics that can be used to inform social intervention to decrease cancer mortality rates. While we found six characteristics associated with cancer death rates, we believe some are not suited to be pursued for social intervention. The characteristics of percent married and percent black are not suitable to pursue to inform social intervention mainly because they showed weak correlations and there are characteristics with stronger correlations available. In addition, promoting marriage may be not be easy to promote and race is something one cannot change. Employment and unemployment percentages do have a strong negative and positive correlations respectively; however, we believe these are secondary to access to good insurance in the form of private vs public insurance. Employed residents likely have access to employee private insurance and unemployed residents would more likely have to rely on public coverage.

This leaves education, access to quality insurance, and poverty as the key characteristics to inform social intervention. Below is a list of possible social interventions based on our findings

Education

Increase the quality of high school education, enable more to qualify for college or both. It is clear that higher education in the form of a college degree had the greatest correlation and a negative one at that. Earning a college degree likely places a person in better position to obtain quality employment that may come with access to employee health coverage. However, there are probably other reasons for this association and may involve the quality of the education. For example, this may be a construct for characteristics such as understanding how to prevent cancer prevention ranging from learning how to live healthy, avoiding carcinogens, understanding the dangers of smoking, and understanding how to utilize the healthcare system. All of these do not require a college degree, hence, increasing the quality of education may be a good social intervention.

Poverty

Poverty was the highest positive correlation to cancer death. There are a lot of reasons to address poverty of people living in the USA and now preventing cancer can be added to that list. Poverty may be associated with no employment or low paying employment. Some of this may be addressed with Education but some may also point to finding ways help people get out of poverty. This may mean temporary financial assistance, finding ways to raise the wages of workers, lower the cost of living, or providing basic financial planning to empower them to save money.

Insurance

Public Insurance has a moderate positive correlation with cancer death rates which is in stark contrast to private insurance moderate negative correlation. Insurance may be viewed as a construct for quality of health care services. It is generally known that healthcare providers do not accept all health care insurance. This is a place to explore the differences between private and public health coverage and identify what might be improved in the public plans.

Analyses to further pursue

Further analysis with this dataset might include utilizing the average annual count of cancer cases variable (avgAnnCount). This variable can be dividing by the population variable (popEst2015) to yield a cancer rate. This can then be used as an outcome for cancer cases using the same exploratory methodology. In addition, if a cancer death to cancer rate ratio can be calculated by dividing the deathRate variable by this calculated cancer rate variable where a ratio closer to zero would indicate a cancers that do not progress to death and a ratio near one would indicate a greater number of cancers progressing to death. Since different types of cancers have different etiologies, it would be prudent to pursue data to identify the different cancer types. A low cancer death to cancer rate ratio may be more dependent on fast spreading cancers versus those that take time to progress.

In conclusion we believe Education, Poverty and Public insurance are the three areas that can inform social intervention. Further analysis on more and different data is warranted to investigate these findings in this exploratory analysis to confirm their validity and reproducibility before identifying the social interventions. These variables can then be used to measure the interventions' outcomes.