Cancer EDA

Vinicio De Sola, Sam Tosaria, Joanna Yu

September 25, 2018

Introduction

As members of the society, Cancer is still one of the most cryptic diseases known to man. In the US, it's the second most common cause of death [1]. Thus, it is important to try to understand how it affects the population and explore if some relationships exist between the incidences of this disease, the related mortality rate and social variables like income, level of education, geography, insurance coverage, etc.

For government agencies, these types of studies are important to help governments allocate resources in fighting these diseases. In this report we analysize the county level data, and our target variable is mortality, expressed as **Death Rate** in the dataset.

We will start the study first by describing the dataset: quality of the data, transformations, and the data cleaning done to it. Second, we will describe the mortality at a high level to understand the overall effect of cancer in the US. Then, we will focus our attention on describing census type: age, race, gender,and population level. After that, we will concentrate our attention on social-economic variables: Income, study levels, insurance coverage, and employment status.

Our second step is to analyze relationships between the target variable and other key variables selected in this study. These will range from geographical maps to show concentrations of the diseases per state, to scatterplots and linear regressions between the key variables and mortality. This will give us a sense of what are the possible ways of alloacting resources in cancer research.

Data Description

The data set includes 30 variables with 1 index variable and 27 other numeric variables that describe the various statistics of each county. The 2 remaining variables (binned income and county name) are factors. There are 3047 observations. 18 variables are expressed in percentages. The variables have been grouped by the following major categories:

Table 1: Data Description

No.	Category	Fields
1	Cancer Incidences	avgAnnCount
2	Income	MedIncome, povertyPercent, binnedInc
3	Population, Age, Gender	popEst2015, MedianAge, MedianAgeMale, MedianAgeFemale
4	Geographic	Geography
5	Household Related	AvgHouseholdSize, PercentMarried, PctMarriedHouseholds
6	Education and Age	PctNoHS18_24, PctHS18_24, PctSomeCol18_24, PctBachDeg18_24, PctHS25_Over, PctBachDeg25_Over
7	Employment Status	PctEmployed16_Over, PctUnemployed16_Over
8	Insurance	PctPrivateCoverage, PctEmpPrivCoverage, PctPublicCoverage
9	Race	PctWhite, PctBlack, PctAsian, PctOtherRace
10	Birth and Death	BirthRate, deathRate

[1] https://www.healthline.com/health/leading-causes-of-death

The data descriptions for some of the variables seem incomplete and certain columns appears incongruous. The following section discusses the data quality issues and data transformations.

Data Transformation and Cleaning

The birth rate in the raw data has a mean of 5.6, meanwhile the death rate has an average of 178.7. The divisors for these fields (per 1,000 or 10,000 or per 100,000 persons) are different. We corrected the divisor for these data points, arriving at the result through intuitive deduction.

The global birth rate as per World Bank is 18.1 births per 1000 people [2]. So an average of 5.6 for US per 1000 population is reasonable. Thus the divisor for birth rates must be one thousand. However the death rate needed to be reassessed and transformed.

Table 2: Summary of key census data

	Mean	Median	Min	Max
Mortality per County	606.34	171	6	38150
Birth Rate	5.64	5.381477705	0	21.32616487
Death Rate	178.66	178.1	59.7	362.8

Firstly, to check the data quality of the average annual cancer occurrences field ('avgAnnCount') we divided it by the estimated population of the county ('popEst2015'), creating a new field called 'CancerPer1000', post data cleaning (discused in the next section) the mean for the field is 5.5 cancer occurrences per 1000 population. This implies the death rate with a mean of 164 must be either per ten or hundred thousand of the population .

However, after dividing by 10 implies a mean or 16.4, it's still too high compared to cancer occurence rate of 5.5 per 1000, but dividing this by 100 would give us a mean of 1.64 deaths per 1000 population. Compared with the cancer occurrence rate of 5.6 per 1000 implying a death rate among cancer patients of ~30% is intuitively reasonable. Thus this field must be transformed by dividing it by 100 to get a consistent, comparable baseline.

Additionally, using a one hundred thousand divisor for the cancer occurrences would over-amplify the effect of the large number of low cancer occurence counties in the data, as illustrated by the low median. Thus we chose the per 1000 of population divisor to counter this effect.

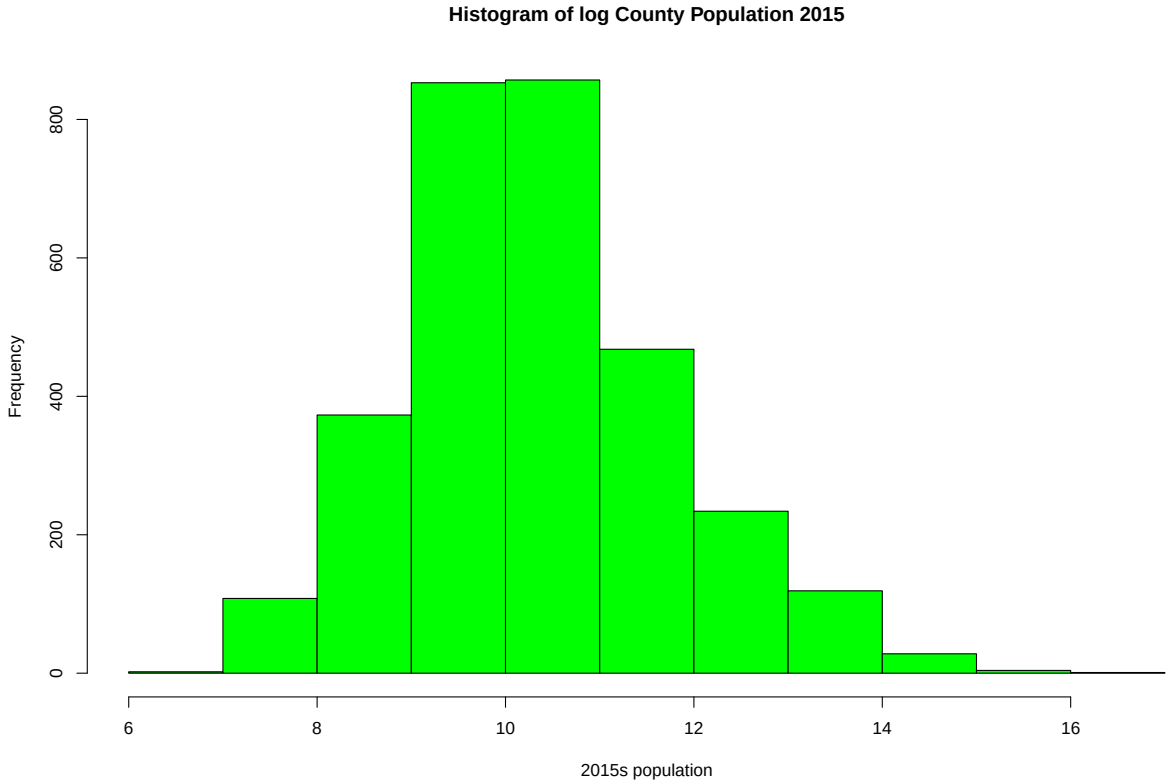
Other varible with problems to use is someBach18-24. The amount of NAs (Not data recorded) is 2285 from 2755 available, thus is not reliable for use, and we discarded.

Having assesed the quality of the data, we proceed to clean it. To clean the data, our first step is to check the influence of population on the data, because we have a huge range in the average mortality per county, and it could be explained by the skewness of the population distribution per county.

Table 3: Summary of County data

	Mean	Median	Min	Max
Population per County 2015	102637.37	26643	827	10170292

[2] https://data.worldbank.org/indicator/SP.DYN.CBRT.IN?view=chart



We plotted the log of the variable, where the amount of skweness in the distribution (Median = 26,643; Mean = 102,637) is quite obvious. Thus, small counties are overrepresented in the dataset. Next, we want to check how age and gender relate to cancer mortality.

Table 4: Data check of Age, Gender and Mortality

avgAnnCount	popEst2015	CancerPer1000	deathRate	MedianAge	MedianAgeMale	MedianAgeFemale
Min. : 6.0	Min. : 827	Min. : 0.9281	Min. :0.597	Min. : 22.30	Min. :22.40	Min. :22.30
1st Qu.: 76.0	1st Qu.: 11684	1st Qu.: 4.8022	1st Qu.:1.612	1st Qu.: 37.70	1st Qu.:36.35	1st Qu.:39.10
Median : 171.0	Median : 26643	Median : 5.6236	Median :1.781	Median : 41.00	Median :39.60	Median :42.40
Mean : 606.3	Mean : 102637	Mean : 23.2443	Mean :1.787	Mean : 45.27	Mean :39.57	Mean :42.15
3rd Qu.: 518.0	3rd Qu.: 68671	3rd Qu.: 6.4874	3rd Qu.:1.952	3rd Qu.: 44.00	3rd Qu.:42.50	3rd Qu.:45.30
Max. :38150.0	Max. :10170292	Max. :2367.5123	Max. :3.628	Max. :624.00	Max. :64.70	Max. :65.70

The total sum of average annual cancer occurrences is 1.847 million. This data observed against the national US population of 330 million implies a 0.56% national cancer occurrence rate. Observing the outlier data in the 'CancerPer1000' field using a threshold of counties showing cancer counts above 1% of the population, there were 198 entries with seemingly dubious average annual cancer occurrences. Of these entries 196 had the exact same data point of 1962.668 producing double and triple digit cancer counts as a percentage of the county population. This data appeared to be an anomaly, thus it was excluded from the analysis. The MedianAge data contained 30 entries with county median age of over 100, logically this appears to be incorrect, so we also omitted these records from the study.

The household size data for 6 counties was very close to zero, this appears to be incorrect data as well, hence we excluded these entries from the analysis as well.

Therefore, our final dataset will not have any of the outliers that we explained before, in the process of cleaning we lost 292 data points, or 9.5% of the raw data. Although the number is not insignificant, it will not skew the Eploratory Analysis.

Table 5: Data check of Household

AvgHouseholdSize	PercentMarried	PctMarriedHouseholds
Min. :0.0221	Min. :23.10	Min. :22.99
1st Qu.:2.3700	1st Qu.:47.75	1st Qu.:47.76
Median :2.5000	Median :52.40	Median :51.67
Mean :2.4797	Mean :51.77	Mean :51.24
3rd Qu.:2.6300	3rd Qu.:56.40	3rd Qu.:55.40
Max. :3.9700	Max. :72.50	Max. :78.08

Univariate Analysis of Key Variables

In this part of the report, we will describe the key variables which may help in policy decisionmaking. In this section we will contune to build upon the analysis started in the previous data cleaning section.

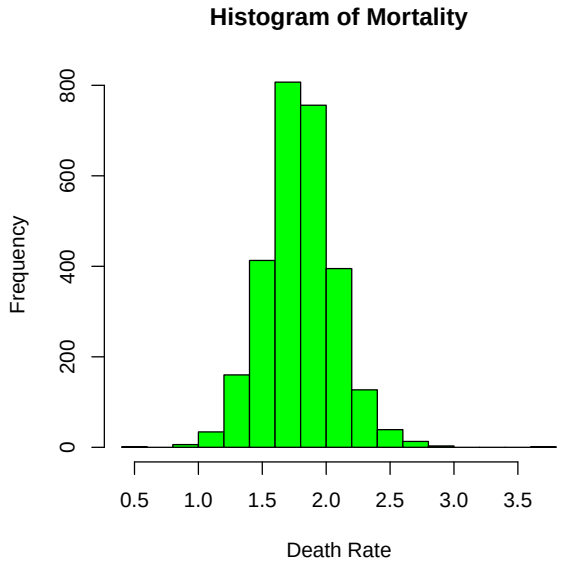
Death Rate (Target variable)

Let's start with the target variable, mortality by observing the summary statistics of this variable from our clean dataset.

Table 6: Mortality of Cancer

	Mean	Median	Sd	Min	Max
Death Rate	1.8	1.791	0.28	0.597	3.628

This distribution seems to be symmetric (Median and Mean are very close to each other). Also, the standard deviation is only 15% of the mean, which mean the distribution should have low levels of kurtosis. Let's graph a histogram and a boxplot of the variable to check the distribuion. As explained in the cleaning section, we are working with deaths per 1000 of population.



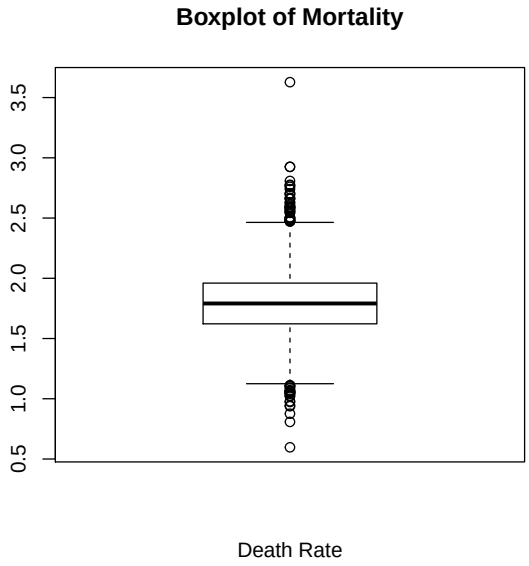


Table 7: Higher Outlier

	County	State
1490	Union County	Florida

As expected from the summary statistics, the distribution is symmetric, almost normal, with low skewness. The box plot doesn't show too many outliers, in fact only one, the maximum value with a mortality rate of 36.28. This outlier is Union County, located in Florida. This outlier is important for us to understand why, because it could mean a potential carcinogen is in that environment, or other exogeneous circumstances.

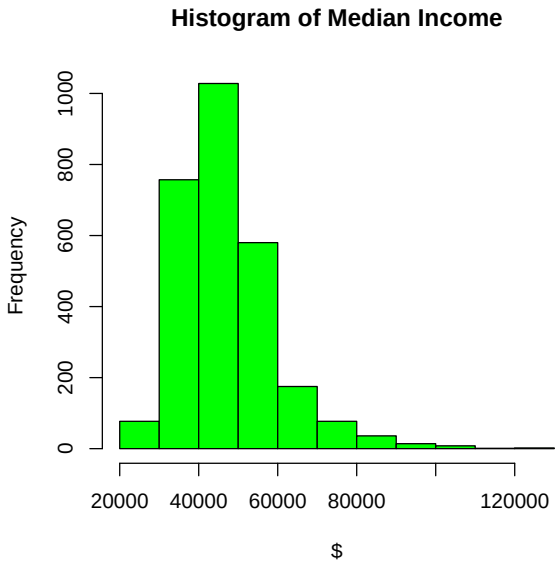
Median Income

Our first independent variable to analyze is Median Income. Now that we have a clean dataset, we can proceed to summarize this variable.

Table 8: Median Income

	Mean	Median	Sd	Min	Max
Median Income	46759.68	44693	12216.86	22640	125635

This is a more skewed variable, with a higher mean that the median. Also, it's quite dispersed, with a high standard deviation of 12K$. This is an important variable in general in terms of socioeconomic impact, thus let's plot the histogram so we can get a better sense of the shape of the distribution.



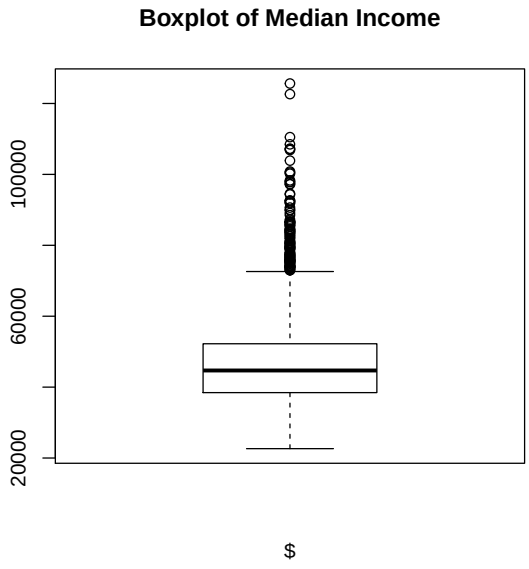


Table 9: Higher Outlier

	County	State
260	Falls Church city	Virginia

Table 10: Lowest Median Income

	County	State
1202	Holmes County	Mississippi

As expected, the distribution is quite skewed to the left, with several outliers in the right tail. The richest county is located in Virginia, called Falls Church City. This type of information is important for Goverment agencies to determine inequality problems and better allocation of welfare resources. The poorest County is locate in Mississippi, and it's called Holmes County.

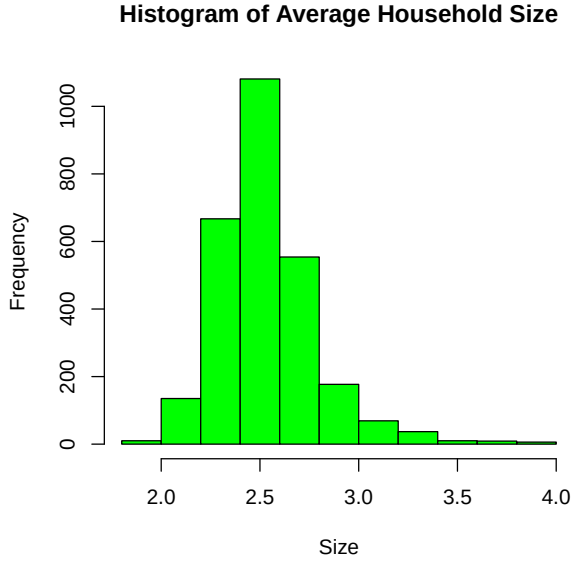
Household Size

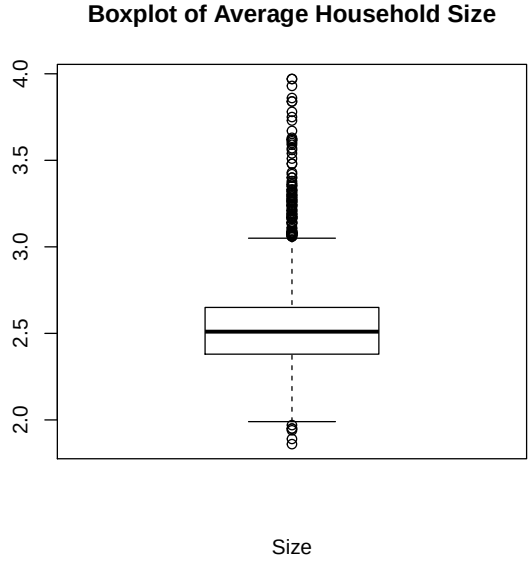
Our second important variable to describe, is the average household size. This is important for density calculation and to check if mortality is related to loneliness, which could shed more light on socia programs those segments of the population.

Table 11: Household Size

	Mean	Median	Sd	Min	Max
Household Size	2.54	2.51	0.25	1.86	3.97

This distribution seems to be symmetric (Median and Mean are very close to each other). Also, the standard deviation is only 10% of the mean, which mean the distribution should have low levels of kurtosis. Let's graph a histogram and a boxplot of the variable to check the distribuion. It is important to take into consideration, this variable is an average, thus it's expected to have less volatility if the Central Limit Theorem is valid for this data.





The histogram showed a different story. The distribution has a small left skewness and a fat right tail. There are two Counties with the largest household size (in average): McKinley County in New Mexico and McKinley County in Alaska. The lowest household size County is Forest County in Pennsylvania.

Table 12: Higher Outlier

	County	State
176	McKinley County	New Mexico
2729	Northwest Arctic Borough	Alaska

Table 13: Lowest Household Size County

	County	State
1723	Forest County	Pennsylvania

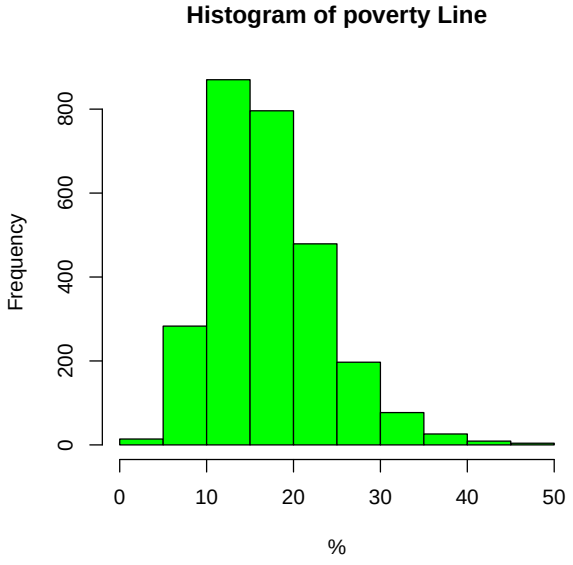
Poverty line

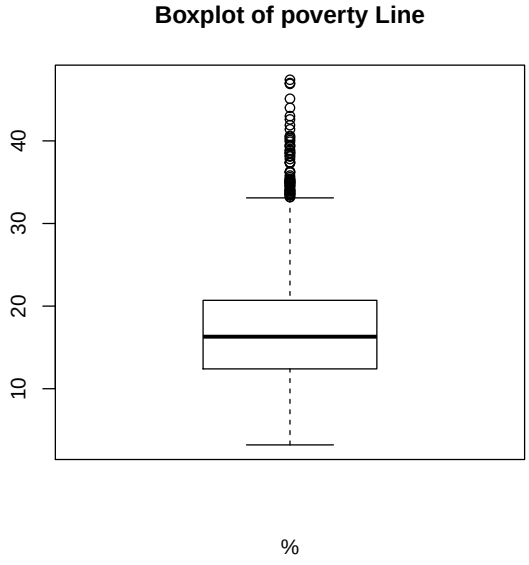
Let's now focus our attention on the poverty line variable. This variable is the percentage of people in the county living below the poverty line. Again, an important socioeconomic variable for allocation of welfare resources. Let's start with the summary statistics.

Table 14: Poverty Line

	Mean	Median	Sd	Min	Max
Poverty Percent	17.2	16.3	6.49	3.2	47.4

it seems to have the same skew as the median income variable. Also, we can identify a quite high maximum value of almost 50%, which means presence of a very poor county in the dataset, compared to a quite low minimum value of 3%, a very rich county. This could help the allocation of resources for decreasing the inequality and possibly the political divisions in the country. Let's take a look at the distribution.





No surprises in the shape of the distribution, it is similar to the distribution of the median income. The poorest county is Todd COunty in South Dakota, and the county with least people below the poverty line is Falls Church City, which also has the highest median income, as mentioned earlier.

Table 15: Poorest County

	County	State
2826	Todd County	South Dakota

Table 16: Richest County

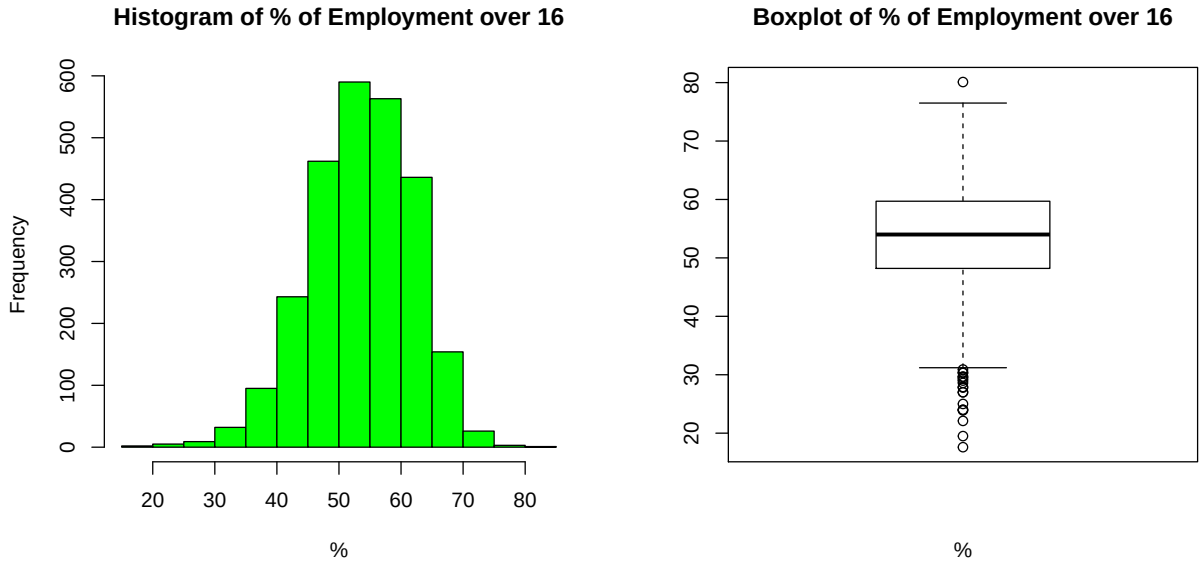
	County	State
260	Falls Church city	Virginia

Employment Status

The last 2 variables that we are going to analyze as key variables are all percentages: employment status of people over 16 and public insurance coverage. These are key socioeconomic and census variables that help localize sources of problems for goverment agencies. In our case, in the next section, we want to explore if these variables have an effect on the cancer related mortality rate. Let's start with employment.

Table 17: Employment Percent

	Mean	Median	Sd	Min	Max
Employment Percent	53.67	54	8.27	17.6	80.1



Is important to state that in this case that the amount of NAs is quite high (134), so this variable have to be treated with care for any of the subsequent explorations.

Insurance Status

Of all the insurance related variables, the one that we care the most about is the coverage by Public agencies or programs, such as Medicaid and Medicare. This is important for tax, political purposes, and of course,

cancer related mortality, since cancer is one of the most cost intensive diseases.

Table 18: Perc of Public Coverage

	Mean	Median	Sd	Min	Max
% of Public Coverage	36.49	36.7	7.91	11.2	65.1

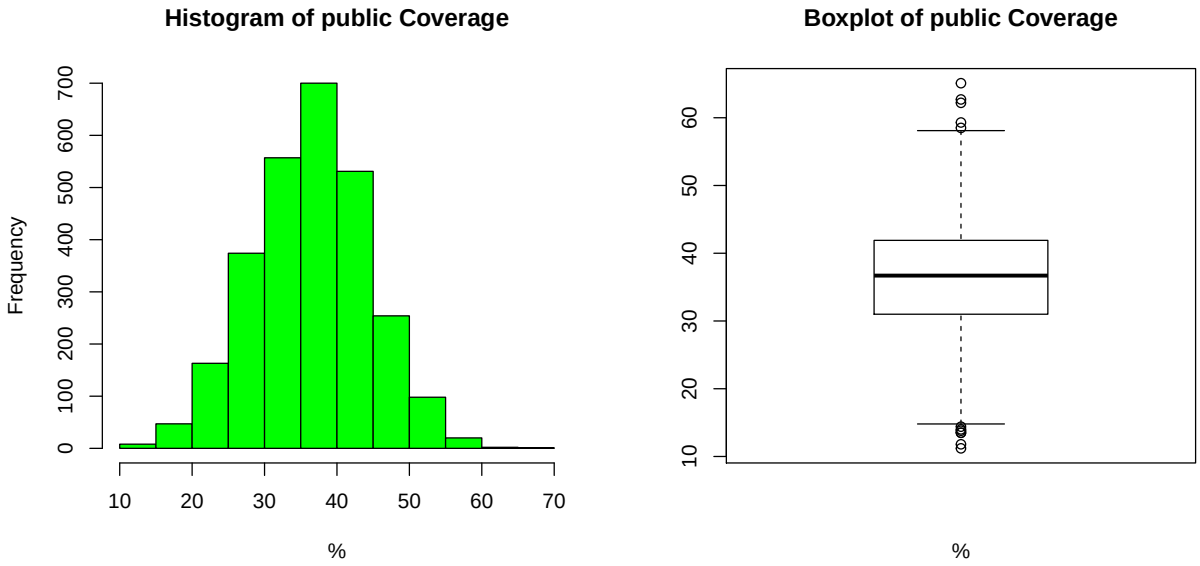


Table 19: Highest Public Coverage

	County	State
1487	Sumter County	Florida

Table 20: Lowest Public Coverage

	County	State
2714	Aleutians West Census Area	Alaska

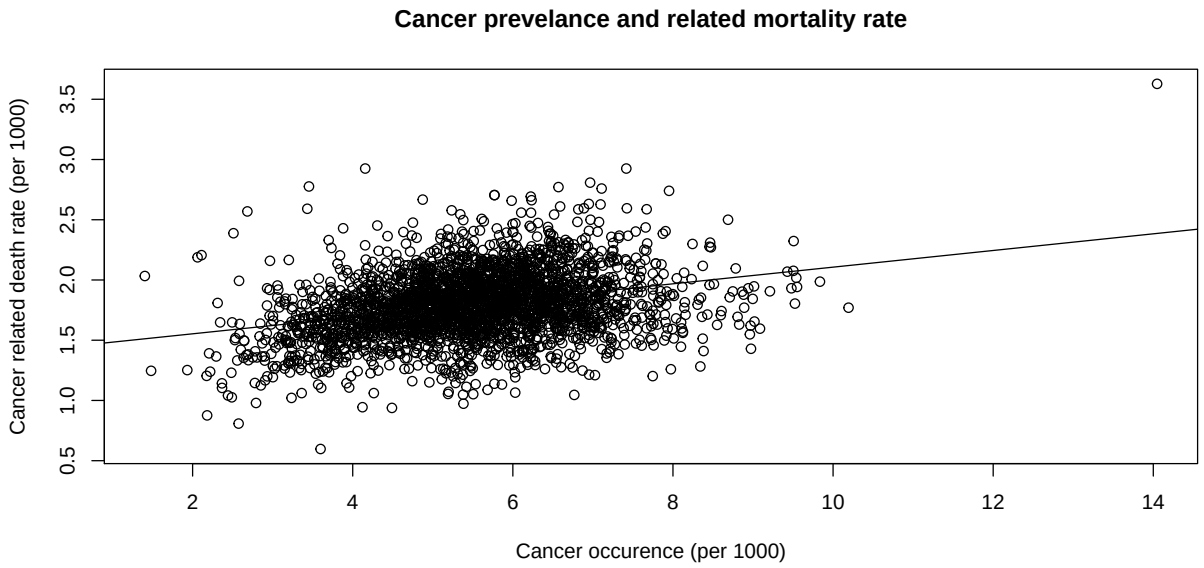
This is quite a symmetric distribution. The county with the highest public insurance coverage is Sumter County in Florida, and the least publically covered county is Aleutians West Census Area in Alaska.

Analysis of Key Relationships & Secundary Effects

After individually describing each of the most important variables, is time to explore the possible relationships that could exist between these variables and our target variable the death rate. As stated before, this is just an exploratory analysis, where we won't model or infer any statistical relationship, but it will help to guide us in further analysis or studies.

Death Rate vs. Cancer Occurrences

Let's first assess our introduced variable (occurrences of cancer), vs cancer mortality. This is a great way to start the analyis, because it will give us a sense of probability of recovery if the person is diagnosed with the disease.

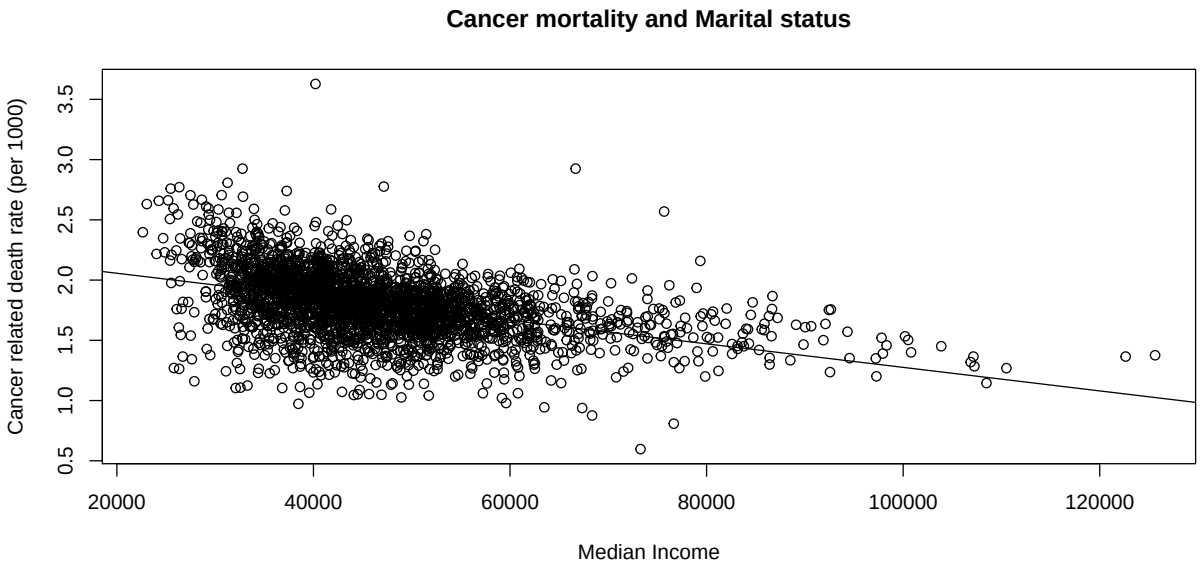


The two are related with a positive correlation of ~30%, as one would expect. As observed in the analysis section there are a confluence of factors influencing the cancer outcomes such as lifestyle, education, income and strong underlying relations between those factors. Thus it is hard to definitively attribute only a few underlying factors to cancer occurence and mortality.

Death Rate vs. Socio Economic Variables

Income

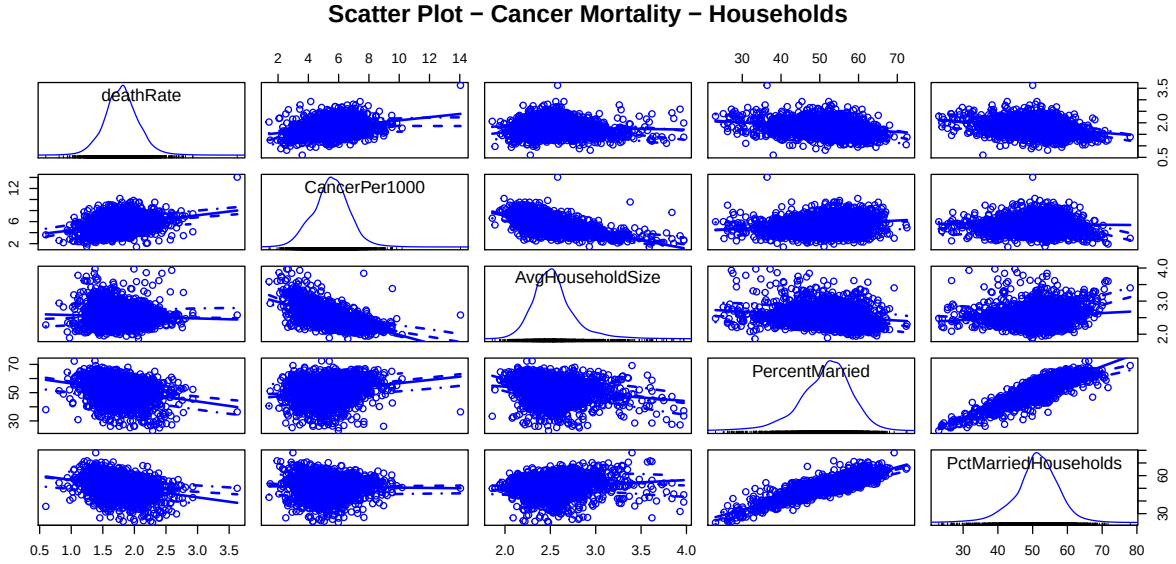
Let's start with Median Income, the most important socioeconomic variable.

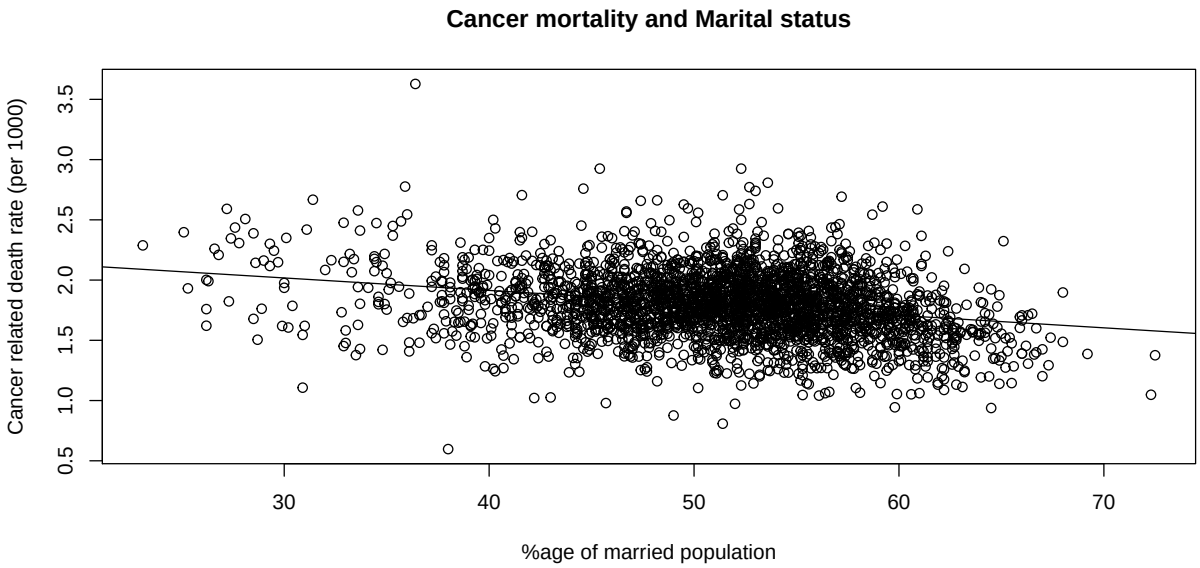


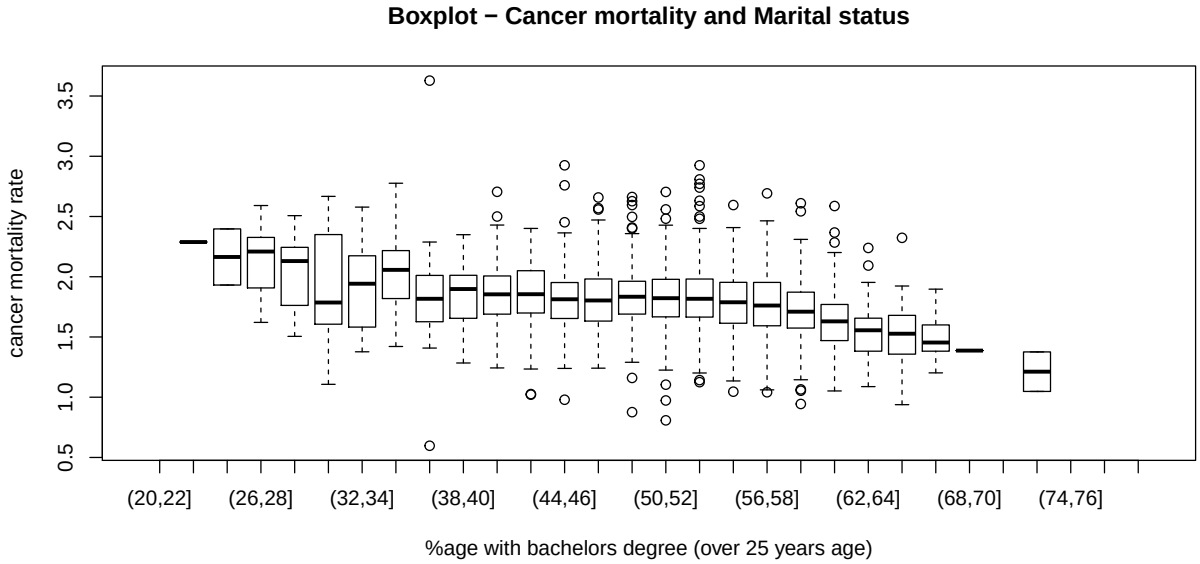
The relationship is quite linear, especially as households with higher income income may have more access to medicines, healtcare, better diet, less stress, etc.

Marital Status and Household Size

Now let's check the relationship between marital status and mortality.







Marital status has a positive correlation with the average occurrence of cancer in the population, though this may be due to the age effect, since counties with more adults and a higher median age would also have more married people.

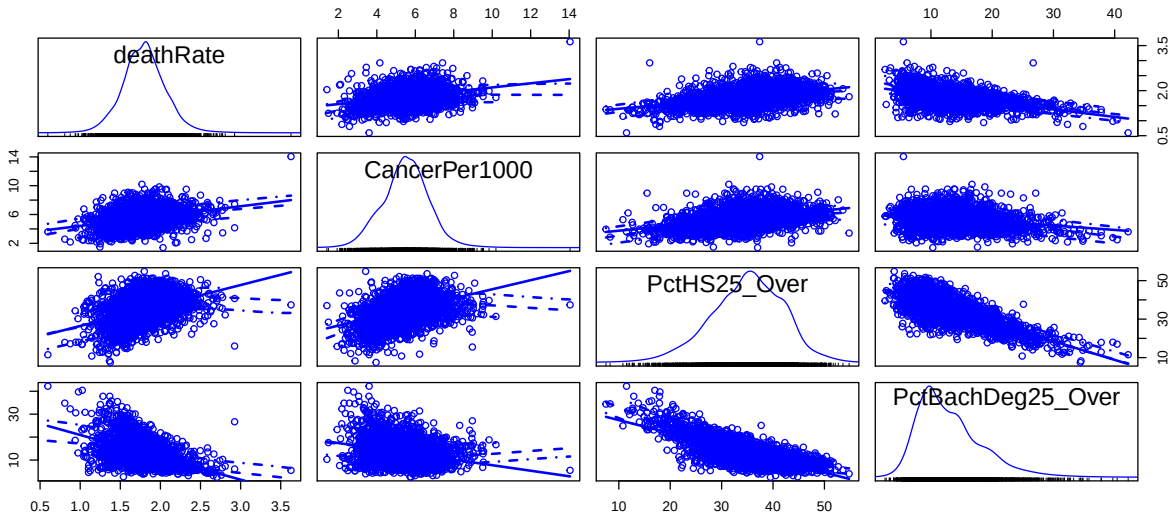
But looking at the death rate and the marital status it appears the counties with higher number of married population exhibit a lower cancer mortality rate with a negative correlation of -0.257

Observing the boxplot of the cancer mortality rate and the marital status of the population, the effect of negative correlation of the death rate with the marital status is even more pronounced for the population between the ages of 58 and 68. This may be due to multiplicity of factors such as better lifestyle, and higher survival rates among those couples. Thus this factor my be skewing the resulting analysis.

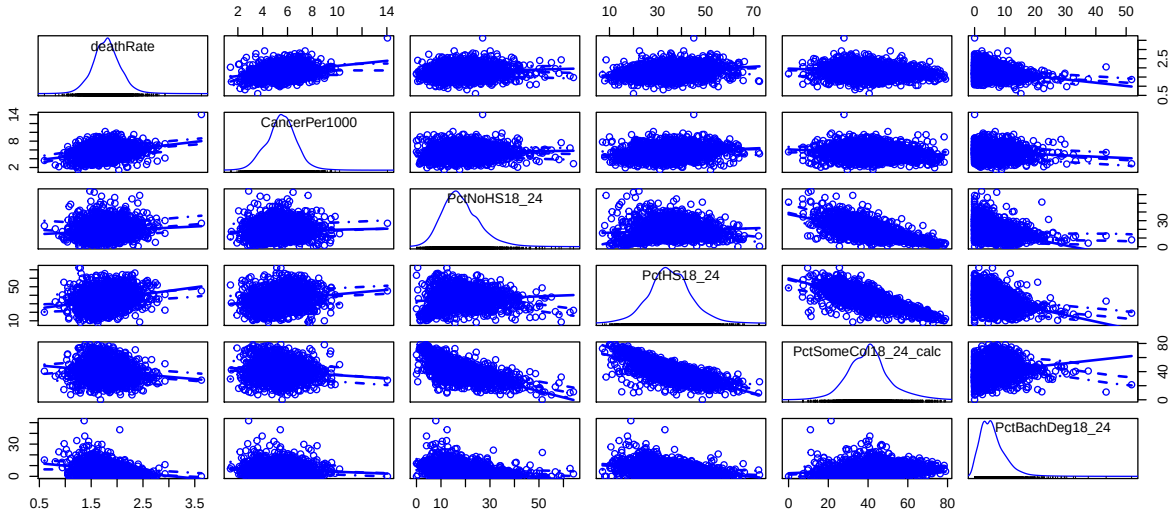
Education

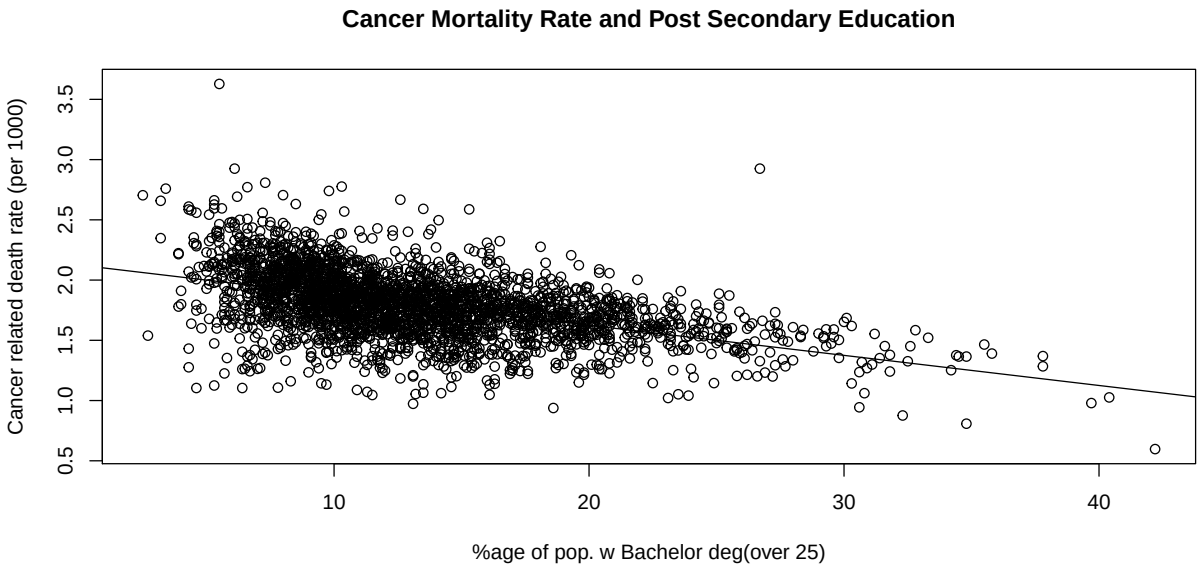
Now let's see the effect of education in mortality. All of these variables may be linked to the first one, median income.

Scatter Plot − Cancer Mortality − Education (over 25)



Scatter Plot − Cancer Mortality − Education (18 − 24)





Plotting the cancer mortality rate alongwith the percentage of population over 25 holding a bachelor's degree illustrates a lower cancer mortality rate among county populations with higher percentage of the population over 25 years of age holding a bachelor's degree. Though this may be due to the same individuals having a higher income level and thus access to better healthcare as well.

Insurance

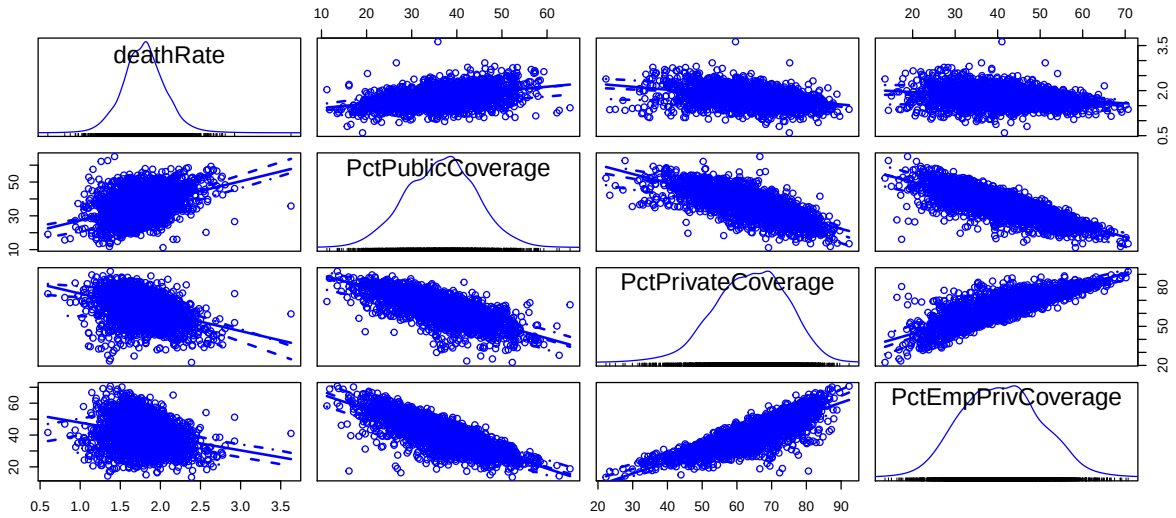
Next we will examine the relationship between the type of insurance people have and cancer mortality by doing a scatterplot matrix. We notice:

1. Cancer mortality is negatively correlated with private insurance coverage (PctPrivateCoverage and PctEmpPrivCoverage).
2. Cancer mortality is positively correlated with public insurance coverage (PctPublicCoverage).
3. As expected, the percentage of people with public insurance is negatively correlated with the percentage of people with private insurance.

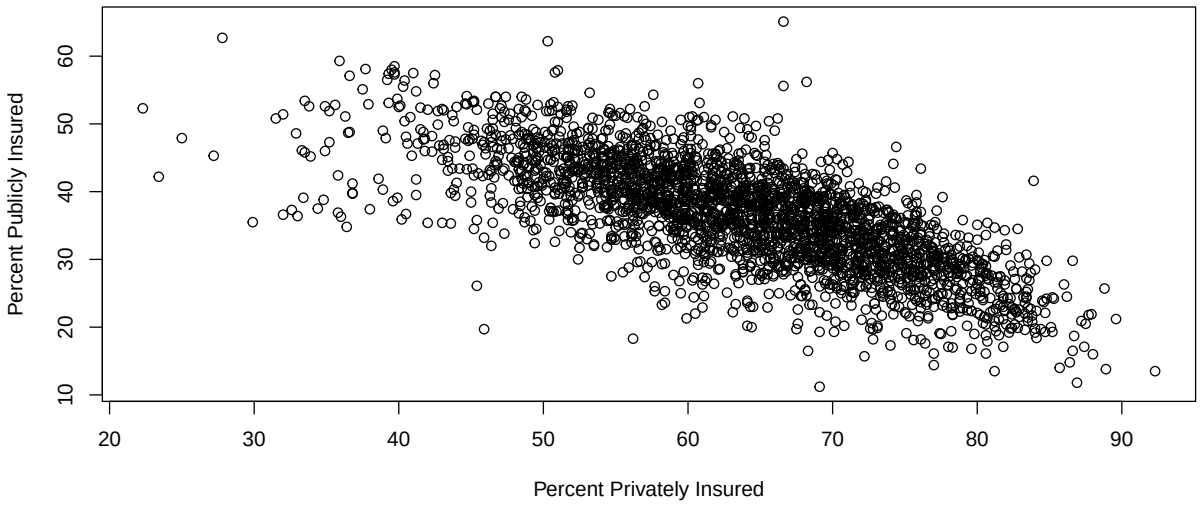
There could be many reasons that cancer mortality is related to insurance. Possible reasons include 1) access to care is often more difficult through public insurance, which could mean less frequent preventative care. This potentially leads to higher mortality if cancer is detected at a later stage. 2) cancer patients may choose to use public insurance programs for their treatment due to the high cost or job loss.

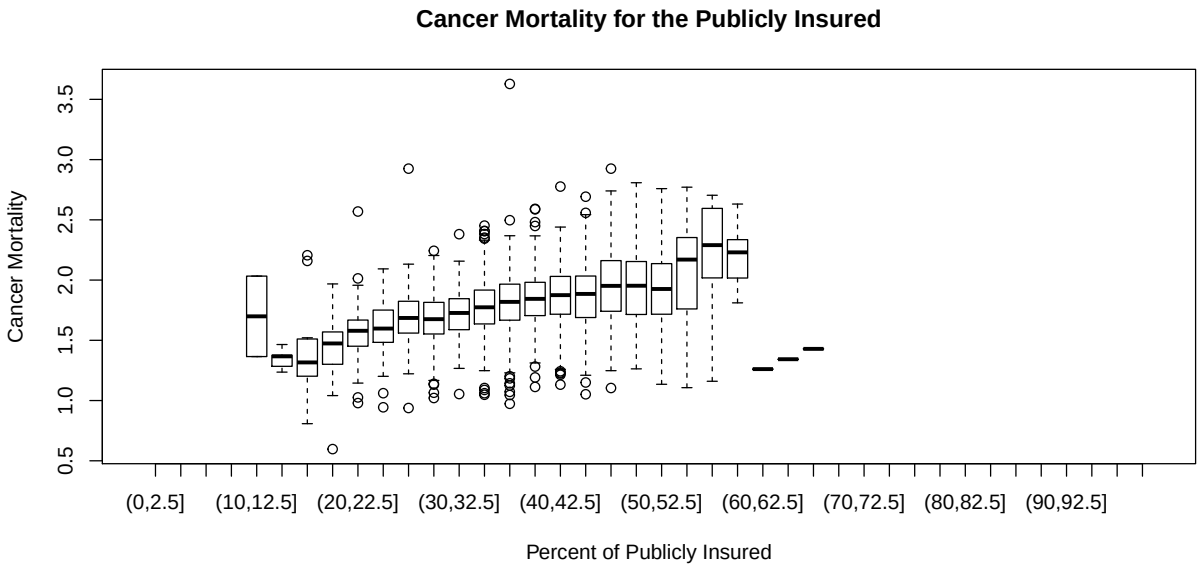
We will now do boxplots to further examine the relationship between cancer mortality and insurance. Each insurance variable will be binned from 0 to 100% by increment of 2.5%

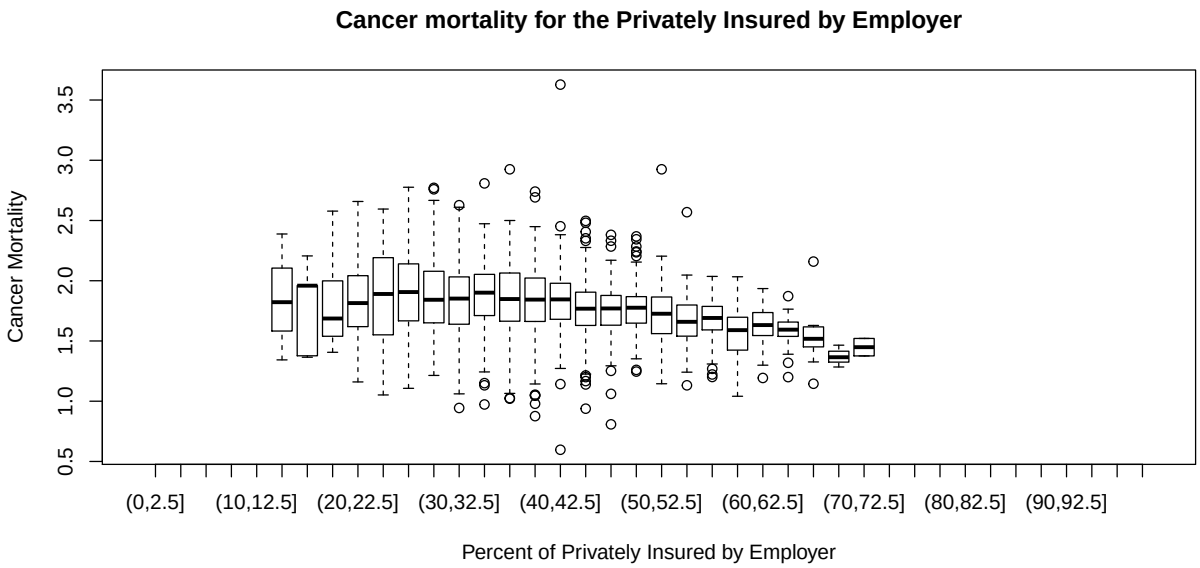
Scatterplot Matrix for Insurance Types



Public Insurance vs. Private Insurance







The boxplots agree with what we observed from the scatterplots. However, there are quite a few outliers in both the publically insured and privately insured.

1. When the percentage of publiclly insured is at its lowest, we would expect relatively low cancer mortality given the positive correlation between the two variables. But cancer mortality is actually high at that point.
2. Likewise, we would expect cancer mortality to be high when the percentage of publiclly insured is at its highest. But cancer mortality is actually at its lowest.
3. 3. When the percentage of privately insured is at its lowest, we would expect relatively high cancer mortality, but it is not the case.

We now attempt to look at the percentage of uninsured to help address these outliers. We assume that if someone does not have public or private insurance, he/she must be uninsured. We create the uninsured

variable by subtracting the percentage of publically insured and privately insured from 100%. But using such method yields many negative values. This means a fairly large part of the population has both public and private insurance. Here is the summary of the uninsured.

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-31.7000  -5.3000  -0.8000  -0.2599   4.0000  34.6000
```

The introduction of the insured variable has helped addressed some of the outliers:

1. When we examine the relationship between the publically insured and uninsured, we can see that when the percentage of publically insured population is at its lowest, the percentage of uninsured is relatively high, which could explain the jump in cancer mortality.
2. When the percentage of publically insured population is at its highest, the percentage of uninsured is negative, meaning people have both public insurance and private insurance. This means they may not be using the public insurance.

Let's now check income with insurance, to study the relationships of secondary effects between variables. At the end, it seems plausible that mortality is more a sum of many variables, rather than an individual factor. We will now look at the income variables. We start by a scatterplot matrix of cancer mortality rate, publically insured population, and income variables. The variable for publically insured population is included because its potential relationship with income variables. As seen in the scatterplots, there is strong positive correlation between the percentage of publically insured and percentage of poverty. The poverty variable may also help explain the relationshiop between cancer mortality and the publically insured population. The positive correlation between cancer mortality and public insurance uptake may be due to the fact that people living with low income or poverty are less likely to be able toafford private insurance. The scatterplot matrix suggests a strong correlation between PctPublicCoverage and povertyPercent (0.65), and a strong negative correlation between PctPublicCoverage and medIncome (-0.75).

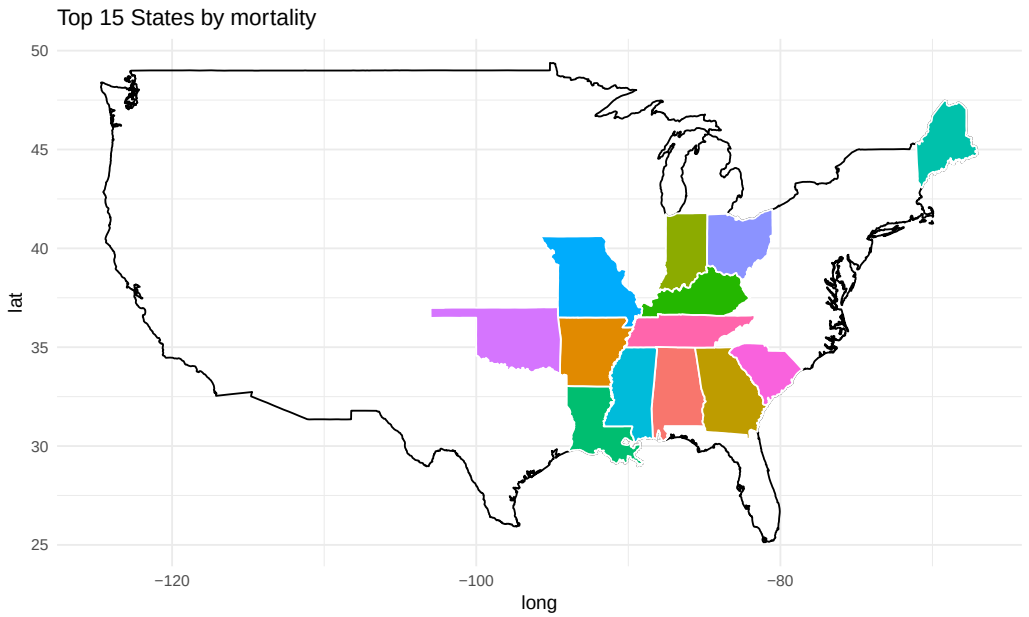
Geography

Our final key relationship to analyze is the geographical data. Where is major concentrations of mortality? This will help any regulatory body to concentrate efforts and resources if one area of the country in particular have a higher mortality rates than others.

```
Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```

Table 21: Top 15 States by Cancer Mortality

	State	deathRate
6	Kentucky	2.150667
9	Mississippi	2.032380
14	Tennessee	2.014817
3	Arkansas	2.000432
7	Louisiana	1.984567
15	West Virginia	1.978302
12	Oklahoma	1.940132
2	Alaska	1.934167
1	Alabama	1.924129
10	Missouri	1.897099
13	South Carolina	1.885800
5	Indiana	1.882011
11	Ohio	1.865965
8	Maine	1.841933
4	Georgia	1.832110



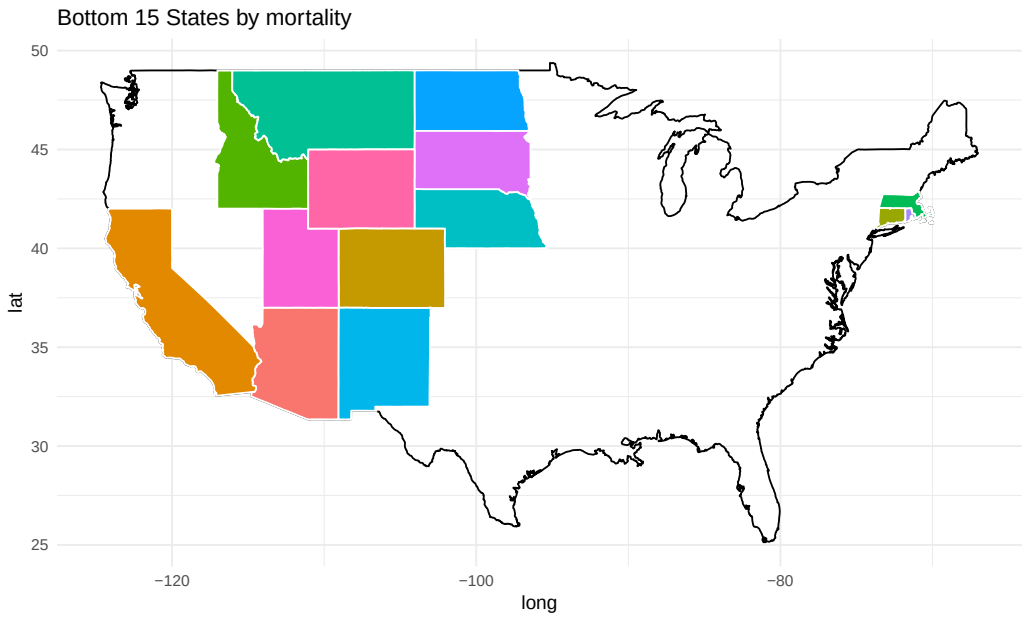
We observe a distinct region of the US where mortality is higher, the South East of the country. This is quite important information for the government and related agencies. This will allow them to assess the allocation of resources and check if there is exogenous environmental variables in that region that could be affecting the cancer mortality rates.

Let's now get the bottom 15 States in terms of mortality.

```
Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```

Table 22: Top 15 States by Cancer Mortality

	State	deathRate
12	Rhode Island	1.659750
7	Massachusetts	1.648000
8	Montana	1.626404
13	South Dakota	1.624035
9	Nebraska	1.603896
11	North Dakota	1.603608
15	Wyoming	1.589870
2	California	1.580965
4	Connecticut	1.577125
10	New Mexico	1.568419
6	Idaho	1.534634
1	Arizona	1.511786
5	Hawaii	1.429750
3	Colorado	1.414068
14	Utah	1.349462



Now the region is also quite clear. The North West Region has the lower cancer mortality. Again, it'll be interisting to see if this can be linked to some environmental or socioeconomic results, because of how concentrated the regions are.

Conclusion

Looking at the education and marital status data the government should encourage more of the population to seek higher education and also encourage more cohabitation among the population as both categories exhibit negative correlation with cancer mortality rates.

From the insurance, employment, and income data perspective, the government may be able to reduce the

cancer mortality rates by 1) improving the public healthcare system; 2) maintaining low unemployment; 3) addressing the issue of poverty.

Since cancer mortality rate is negatively correlated with the percentage of population which is privately insured, and positively correlated with the percentage of population that is publically insured, it is possible that the cancer patients can get better quality of care or better access to care from private insurance. The government can work on improving the public healthcare system to help reduce cancer mortality. In addition the governments should also considering reforming the insurance industry and incentivizing the private insurance companies to offer broader insurance programs, such that people with higher healthcare needs or sickness are able to access healthcare and are able to have more choices rather than be force to only use only the public insurance options provided by the governments.

Since many employers provide private insurance, maintaining low unemployment rate would also help maintain a higher rate of private insurance coverage among the population, alleviating the burden from the public healthcare system.

People living in poverty are more likely to be susceptible to cancer mortality because they may not have the time or resources to access good preventive care. The government can create more outreach programs aimed at offering convenient preventive services at local communities. Improved and continuous wellness monitoring can lead to better treatment when cancer cab be detected early.