Cancer EDA

Introduction

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

Number of Variables: 30
Number of Observations: 3047

Variables:

This dataset contains variables describing county, region, population, birthrate, race, marital status, insurance coverage, income status, and education.

```
##  [1] "X"                     "avgAnnCount"           "medIncome"
##  [4] "popEst2015"            "povertyPercent"        "binnedInc"
##  [7] "MedianAge"             "MedianAgeMale"         "MedianAgeFemale"
## [10] "Geography"             "AvgHouseholdSize"      "PercentMarried"
## [13] "PctNoHS18_24"          "PctHS18_24"            "PctSomeCol18_24"
## [16] "PctBachDeg18_24"       "PctHS25_Over"          "PctBachDeg25_Over"
## [19] "PctEmployed16_Over"    "PctUnemployed16_Over" "PctPrivateCoverage"
## [22] "PctEmpPrivCoverage"    "PctPublicCoverage"     "PctWhite"
## [25] "PctBlack"              "PctAsian"              "PctOtherRace"
## [28] "PctMarriedHouseholds" "BirthRate"             "deathRate"
```

Variable clarification and assumption

PctPrivateCoverage: “Percentage of the population with private insurance coverage”
avgAnnCount: “2009-2013 mean incidences per county WHAT DOES THIS MEAN????”
povertyPercent: “Percent of population below poverty line”
popEst2015: “Estimated population by county 2015”
PctPublicCoverage: “Percentage of the population with public insurance coverage”
deathRate: “Number of deaths attributed to cancer”
binnedInc: “Income groups????” medianAge: “We removed all median ages above 100”

Data Quality

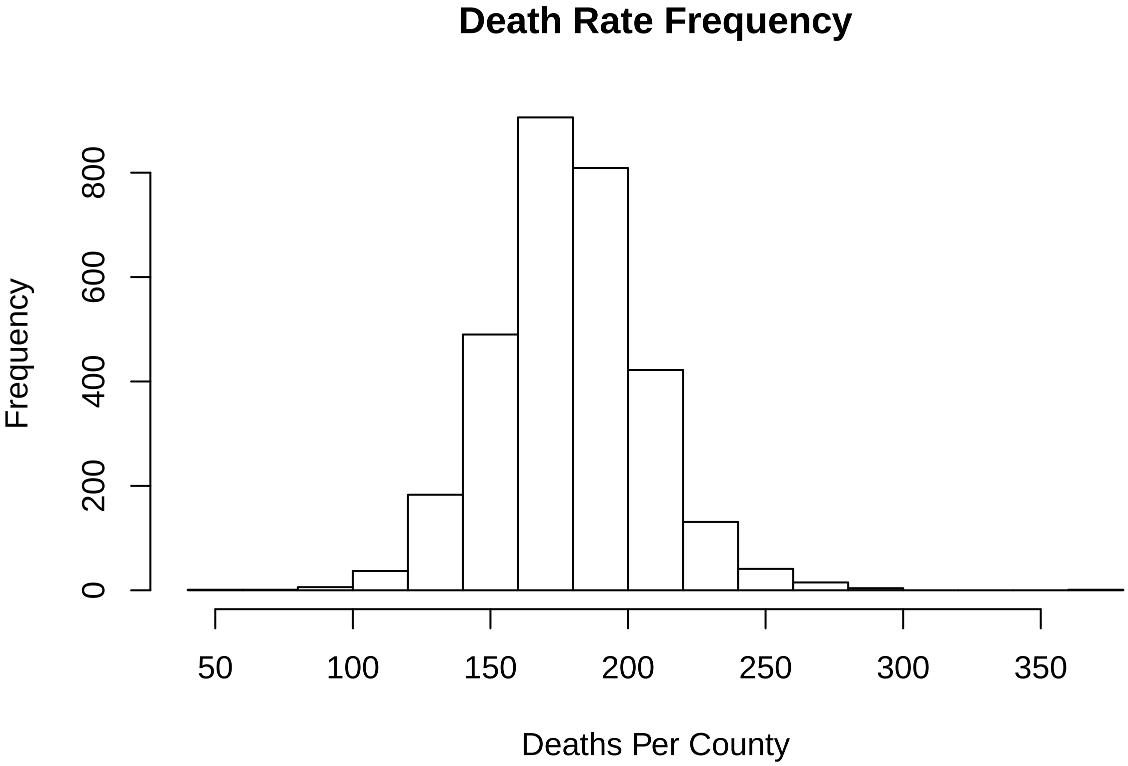
1) The sample size seems to be large enough to get valuable insight.
2) The data seems to be collected in different number formats, even for the same columns. Some have integers, some have floats with one decimal, others many decimals.
3) Seems to be a number of obeservations that are NA of 18-24 with some college, 2285 to be exact.

Univariate Analysis of Key Variables

The key variables that we focused on are in groups related to the variable deathRate:

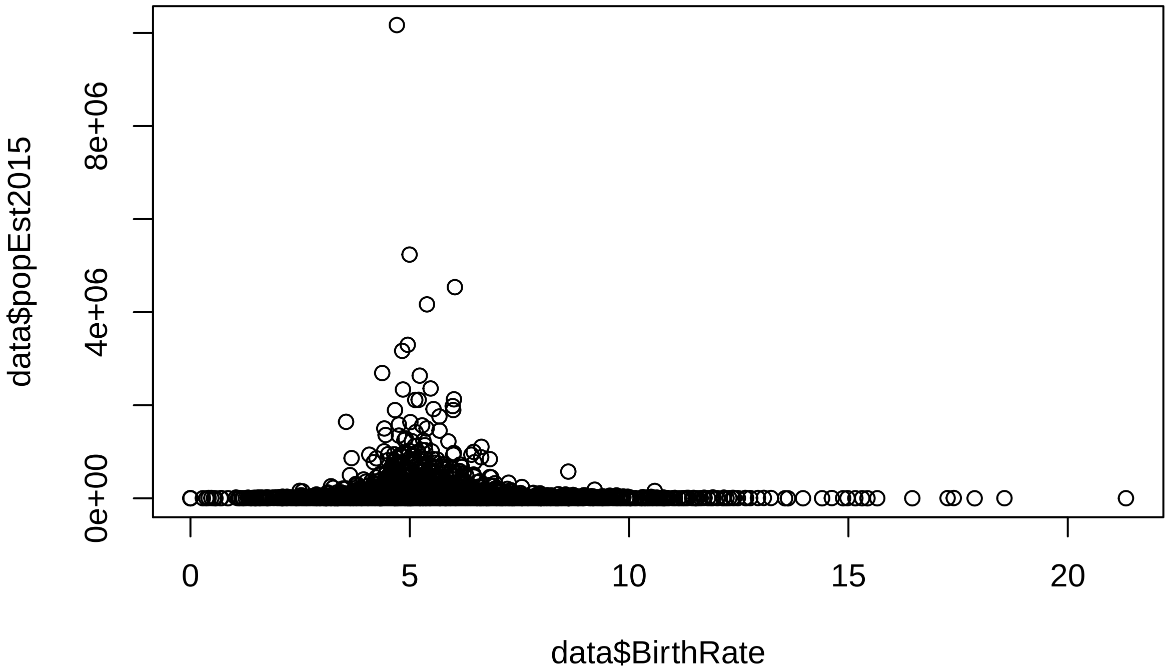
Death Rate

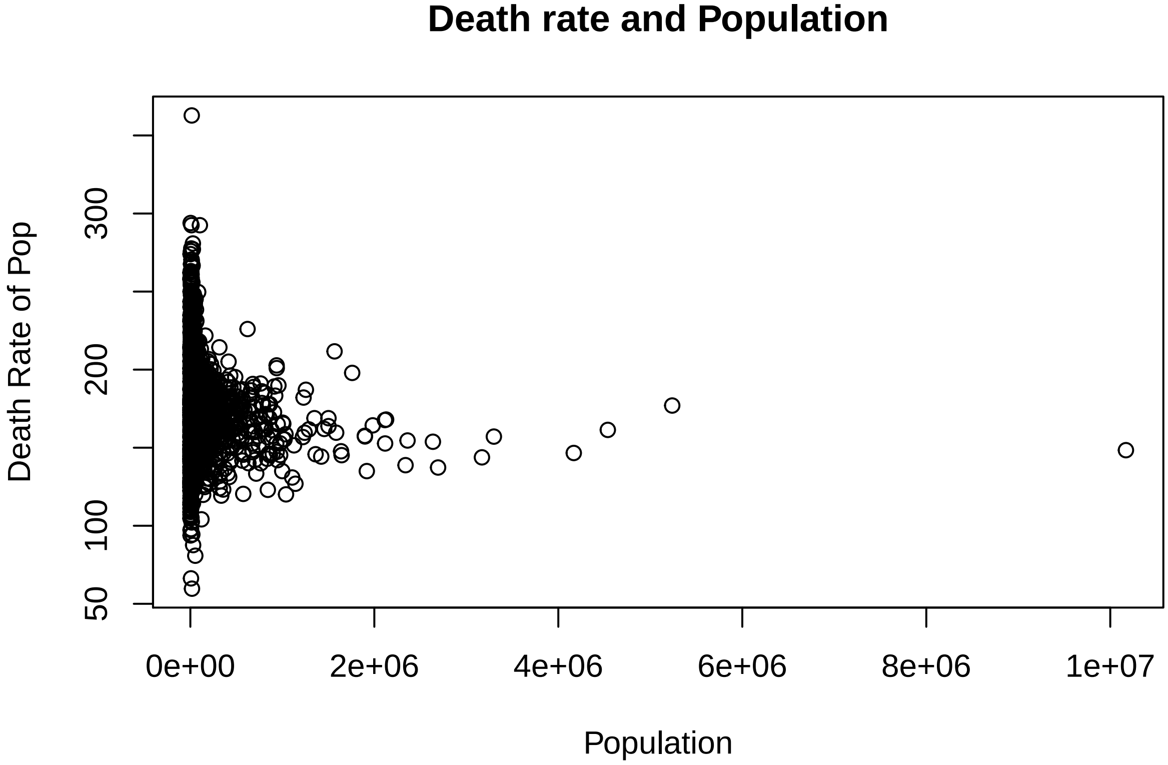
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    59.7   161.2   178.1   178.7   195.2   362.8
```

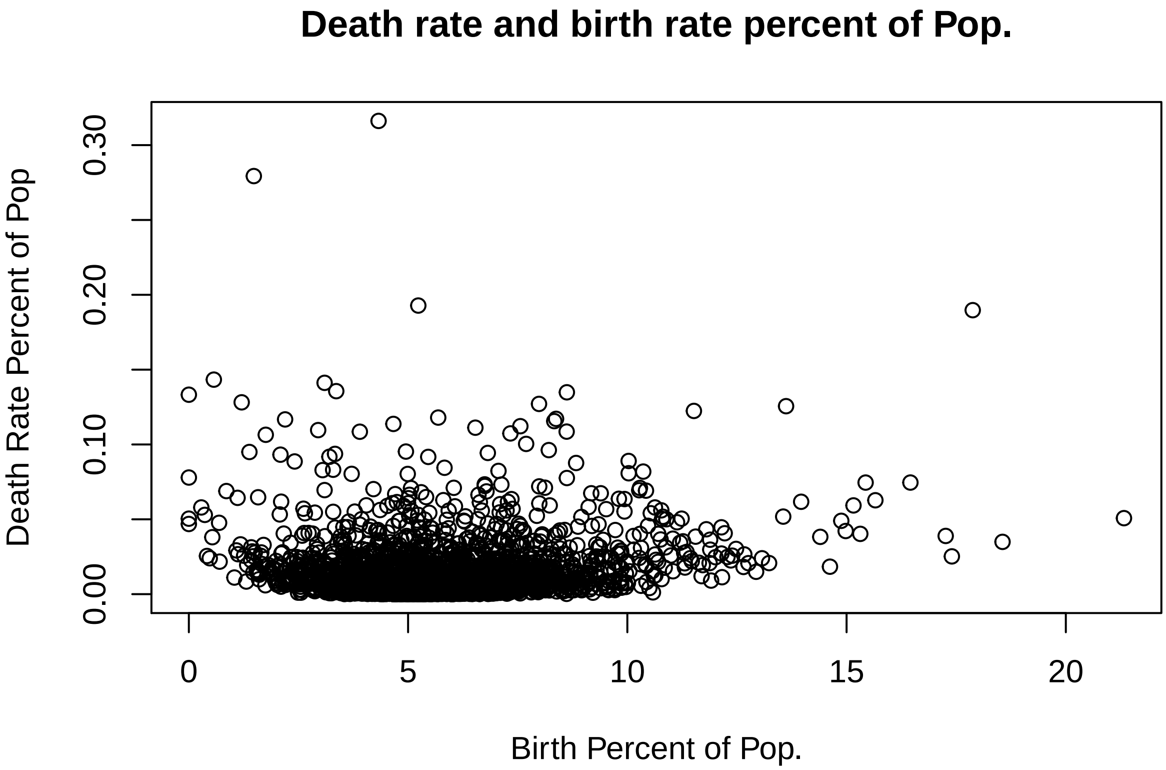


The avg deathrate of cancer is between 150-200

Population: popEst2015, AvgHouseholdSize, PercentMarried, Geography, avgAnnCount, BirthRate, binnedInc

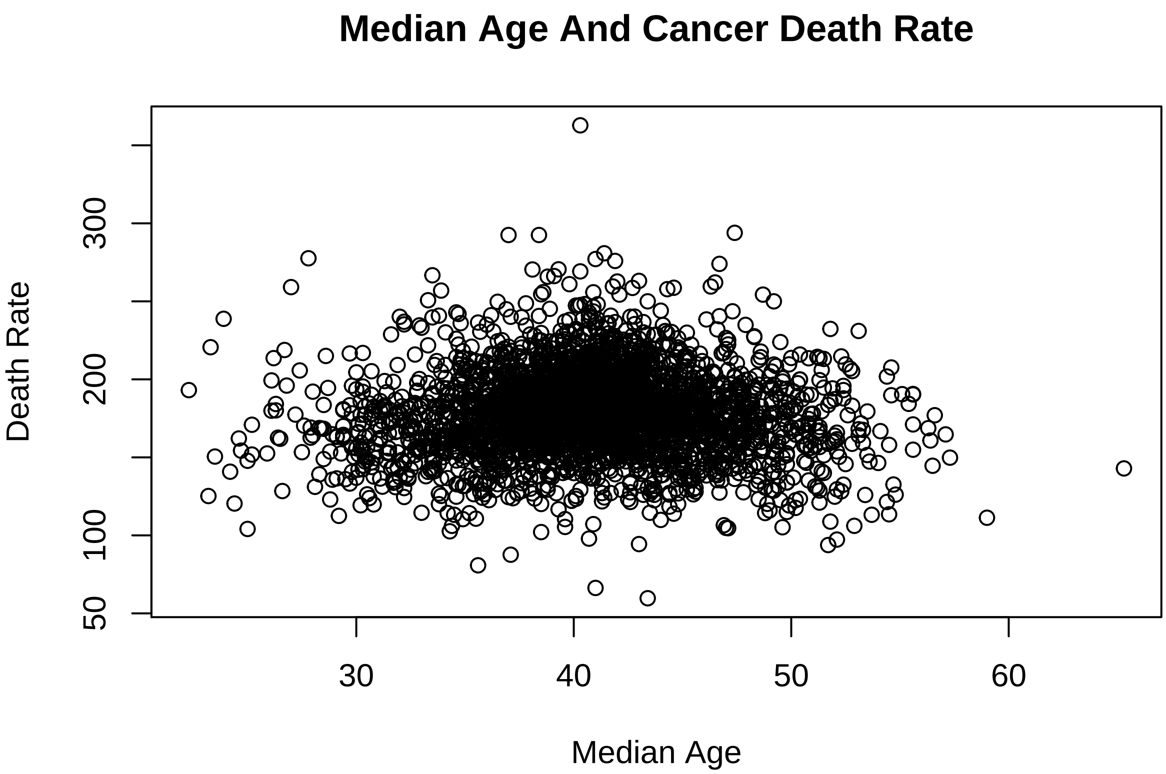


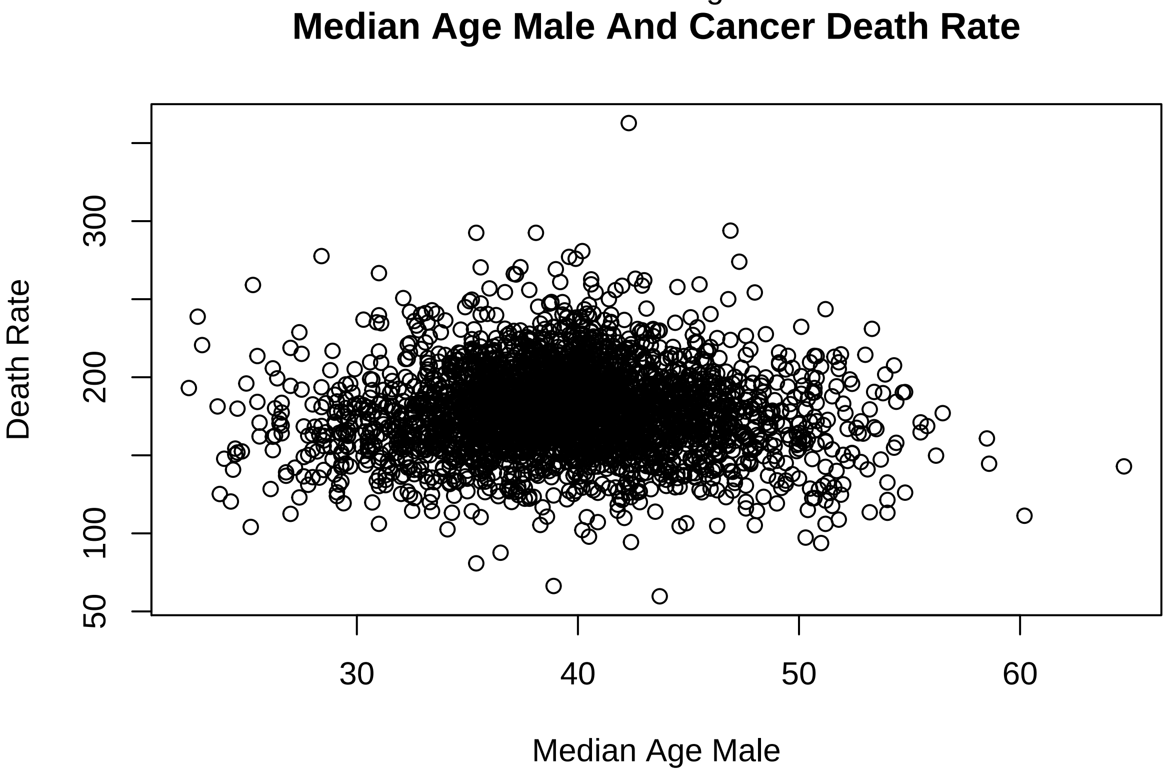


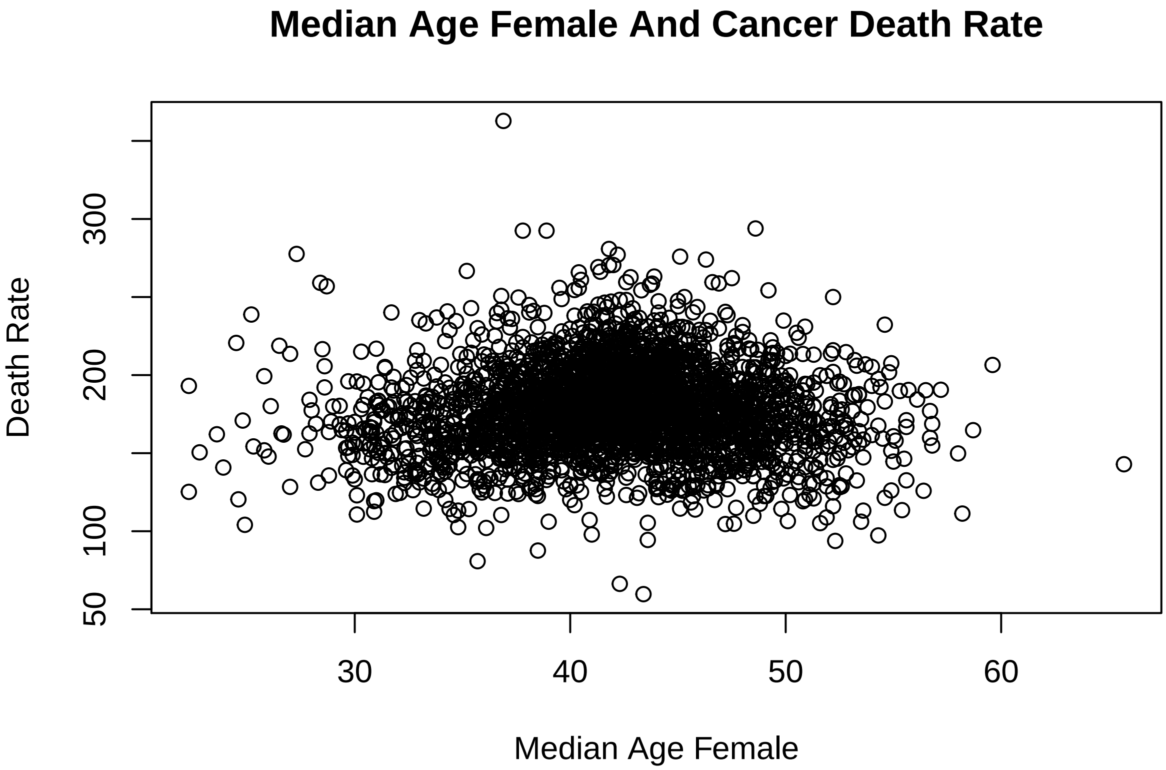


Conclusion on Population variables: - It doesn't seem that high birth rate or population correlates to higher cancer mortality. - We removed avg household size, percent married, geography, angAnnCount, and binned income from analysis due to perceieved irrelevance.

Age: MedianAge, MedianAgeMale, MedianAgeFemale

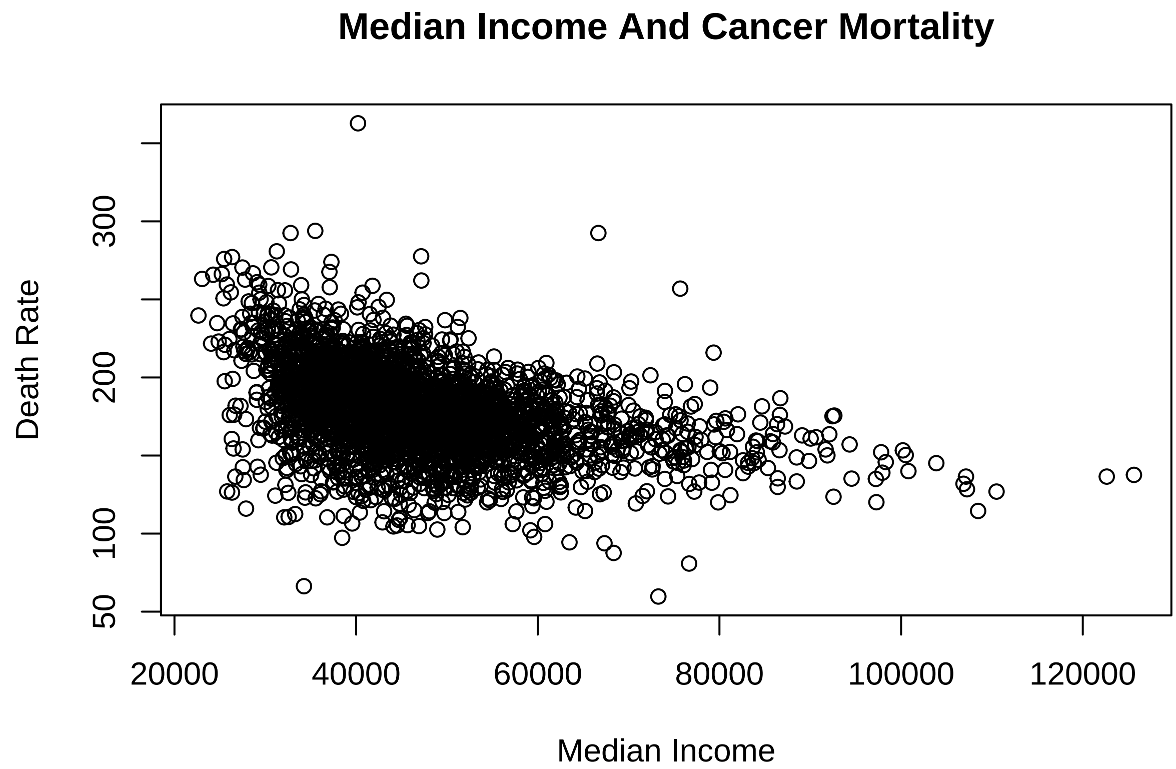




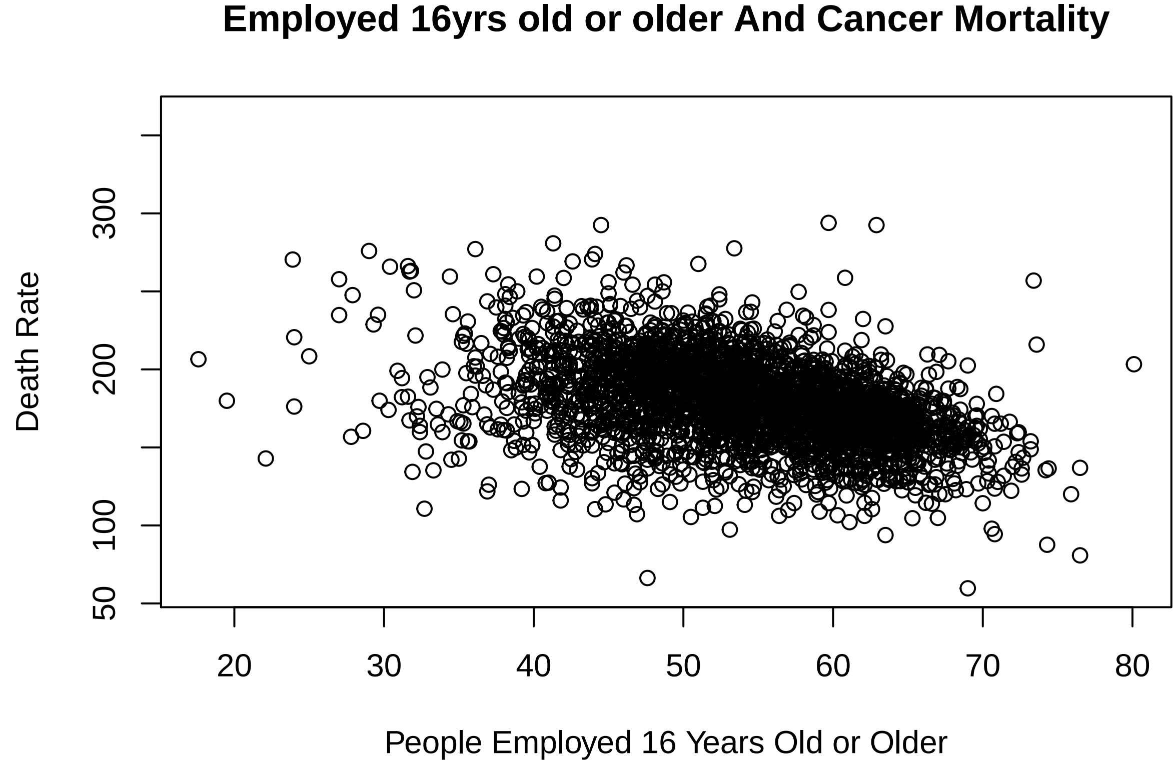


Conclusion of Age Variables: - We removed all median ages above 100 due to some anomalies of median age 300+. - There seems to be a large cancer mortality rate between the 30-50 years of age. - Women seem to group just above 40 and men just under 40 with county deathrates.

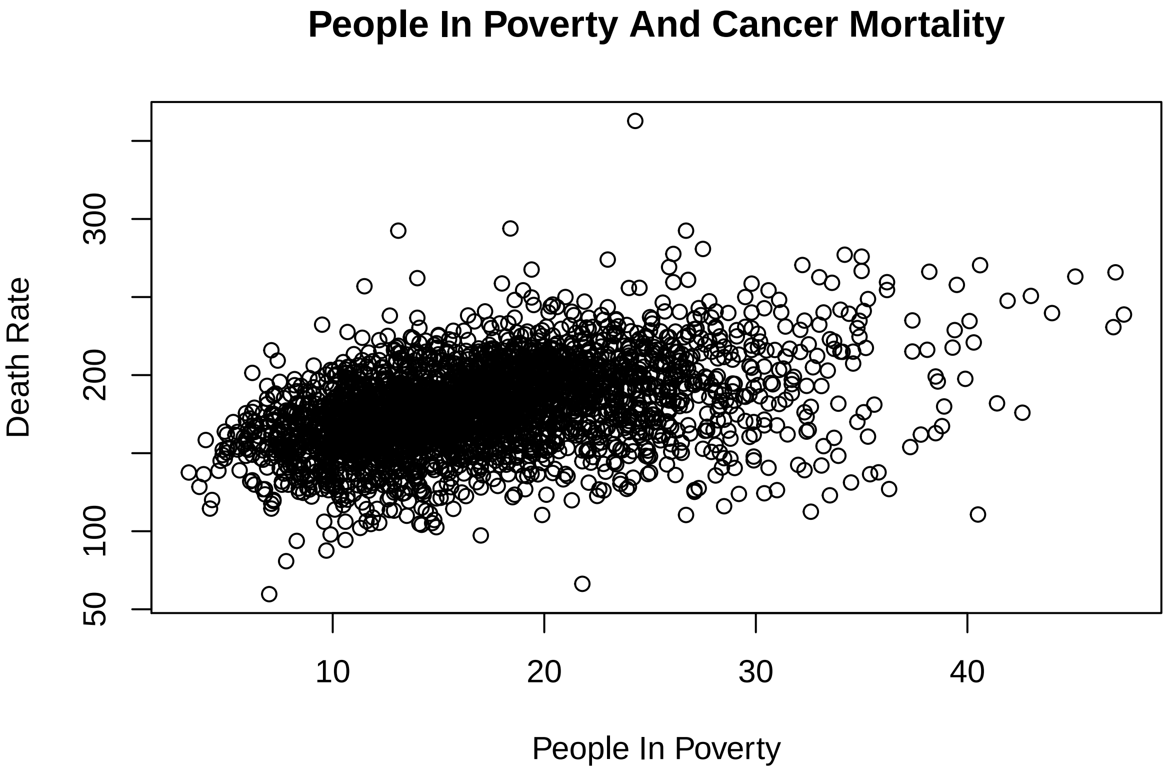
Income: medIncome, povertyPercent, binnedInc, PctEmployed16_Over, PctUnemployed16_Over

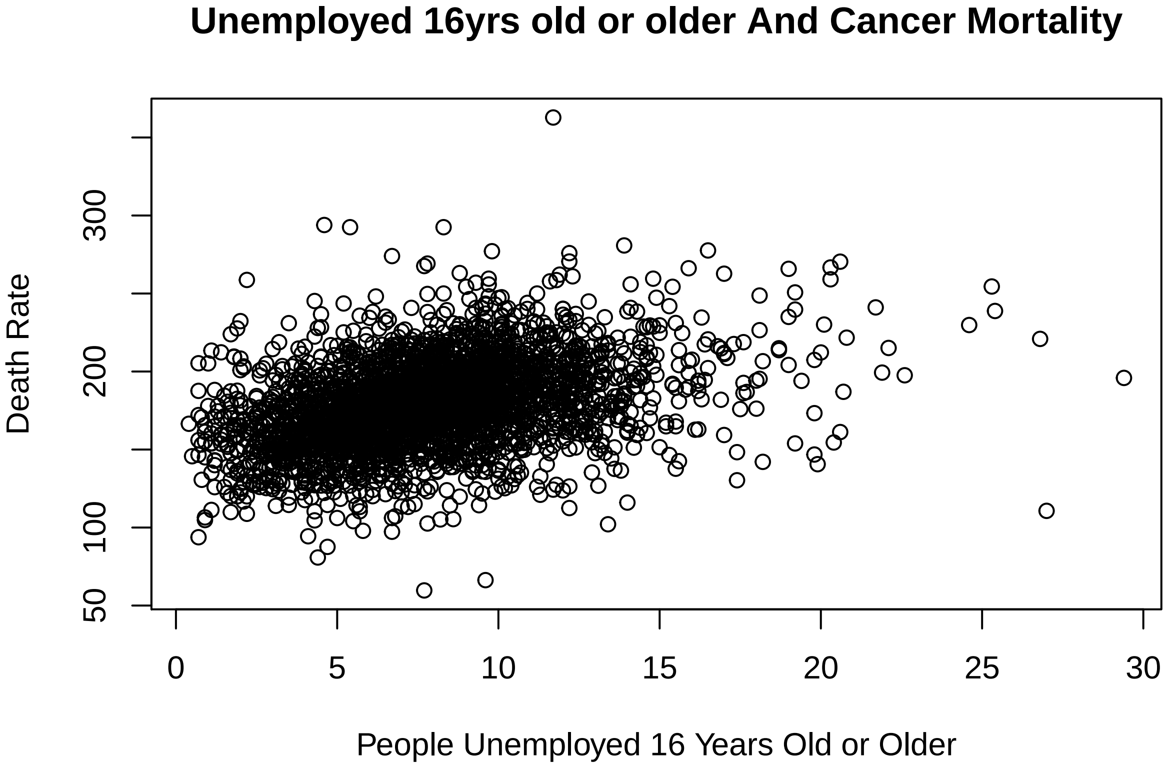


Median Income And Cancer Mortality



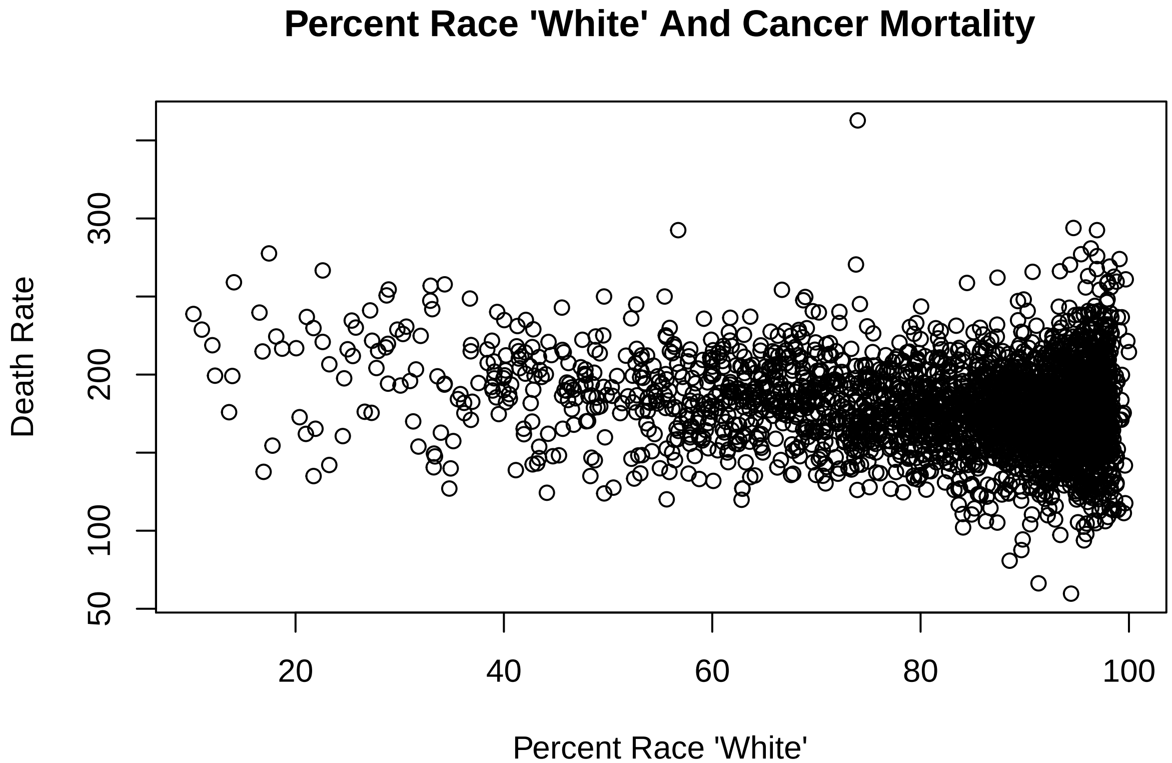
Employed 16yrs old or older And Cancer Mortality

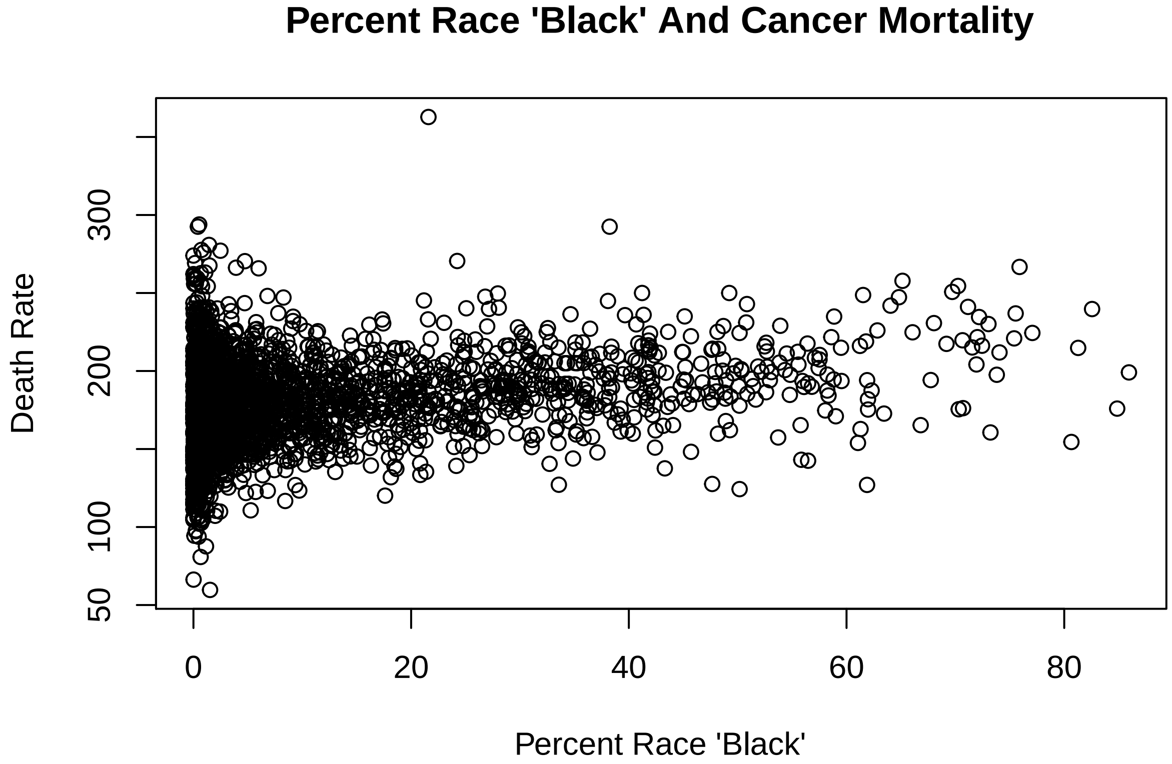


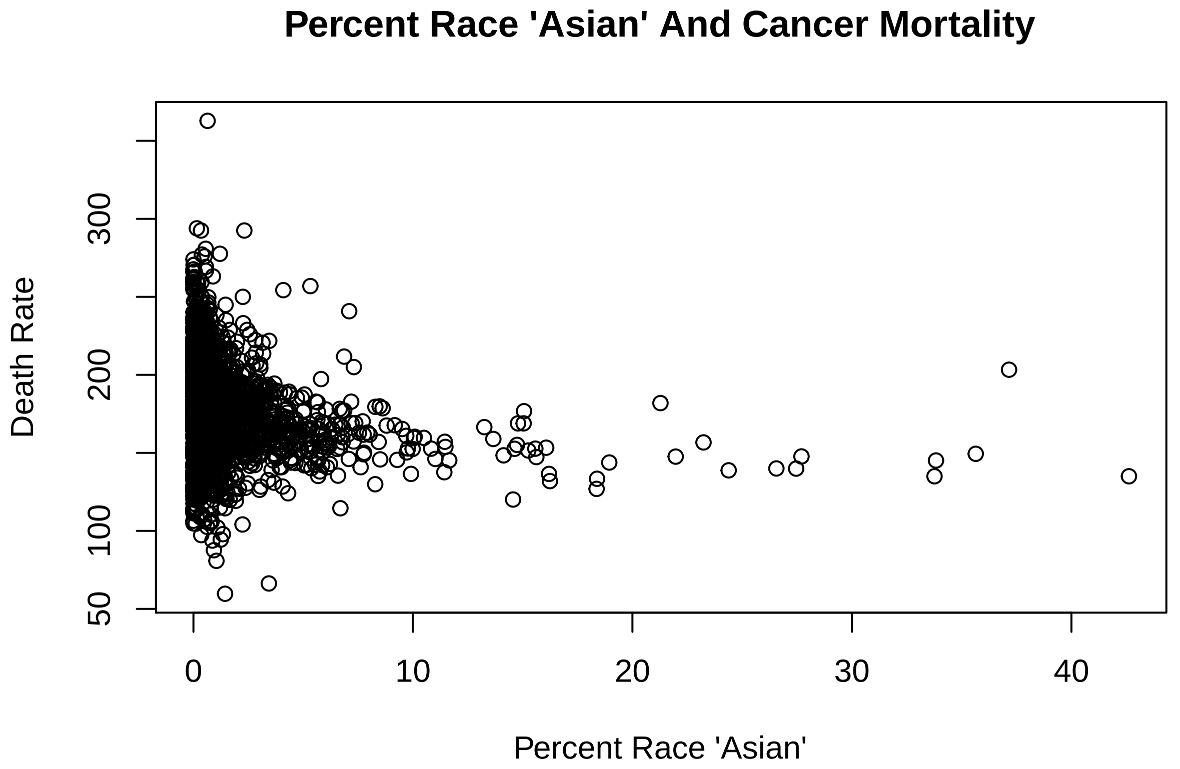


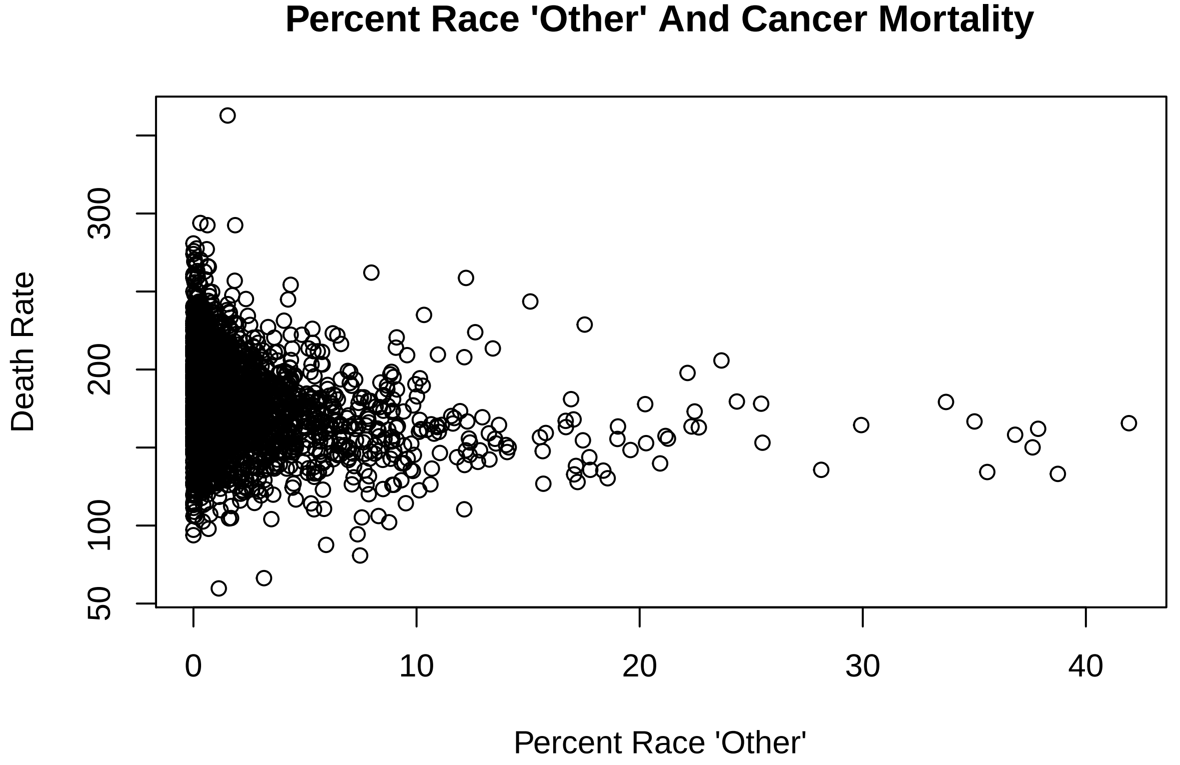
Conclusion on Income Variables: - The strongest correlation yet - As poverty and unemployment goes up, so does cancer mortality - As median income and employment rise, cancer mortality decreases

Race: PctWhite, PctBlack, PctAsian, PctOtherRace



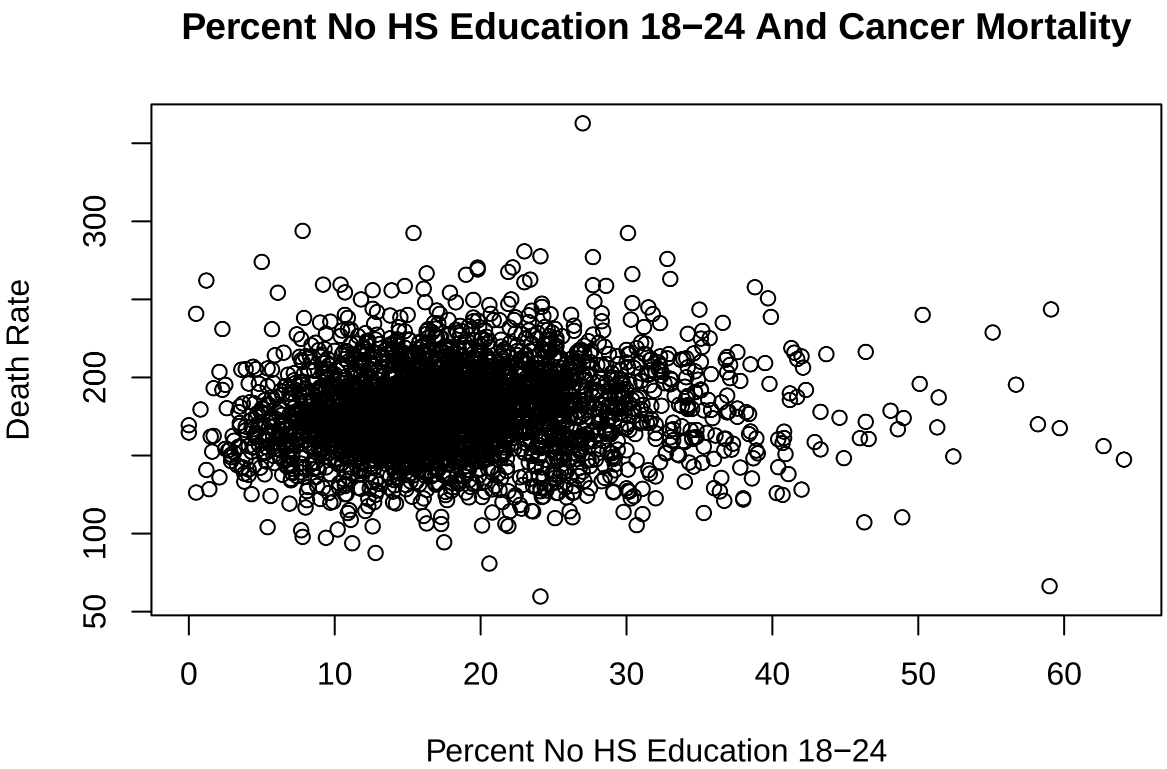


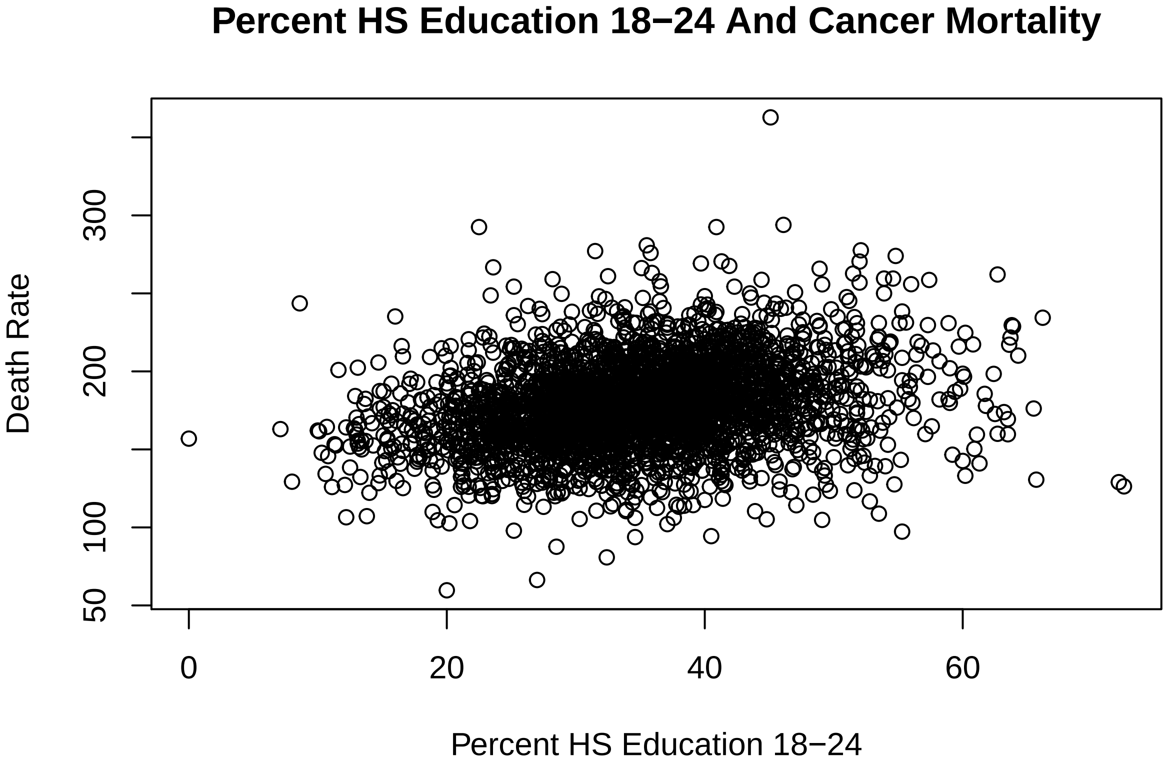




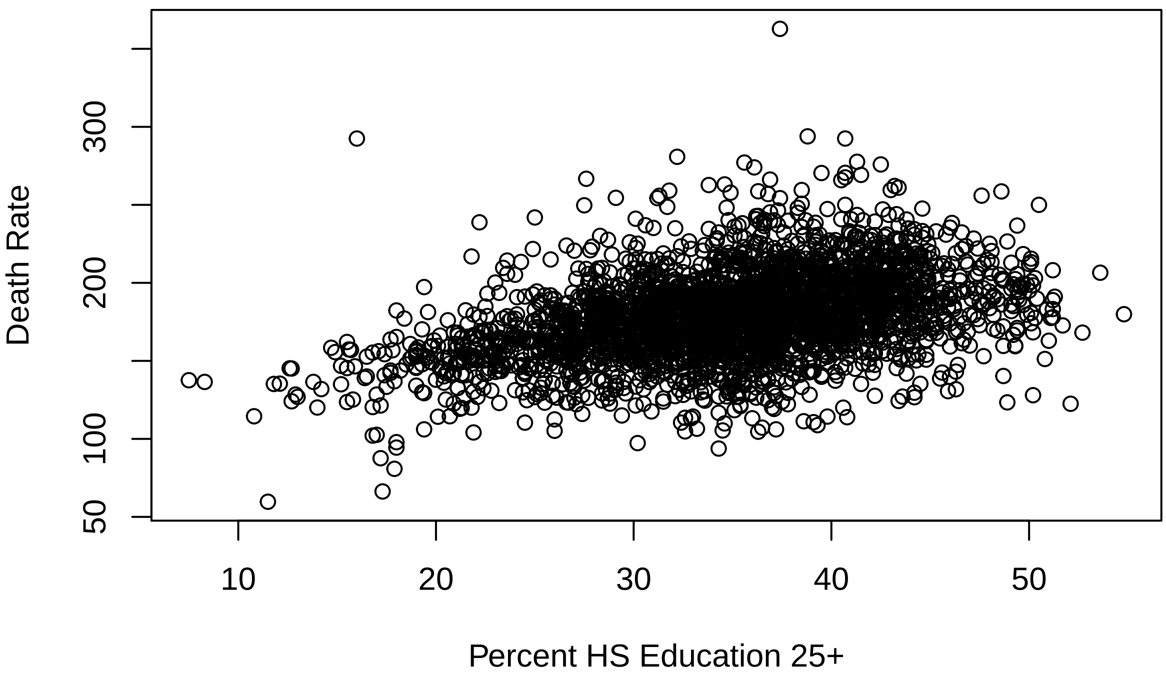
Conclusion on Race Variables: - It seems that many of the counties surveyed were a moajority race 'White' - The death rate seemed to hover around its avg for every race, no major correlation detected

Education: **PctNoHS18_24, PctHS18_24, PctHS25_Over, PctSomeCol18_24, PctBachDeg18_24, PctBachDeg25_Over**

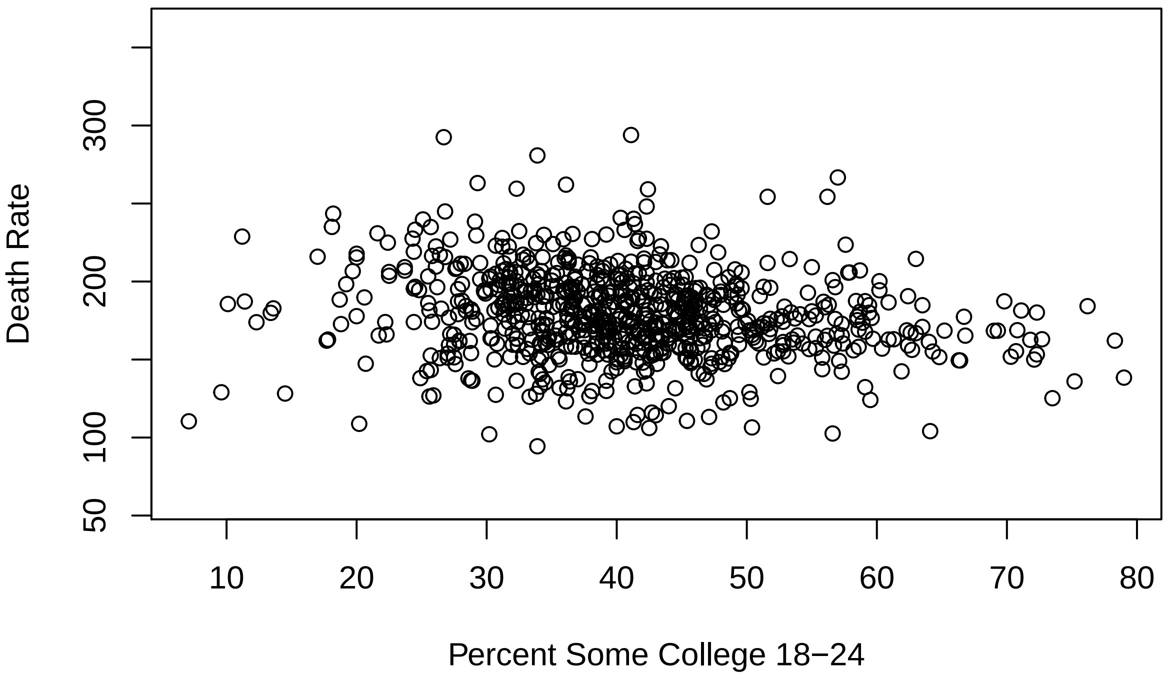


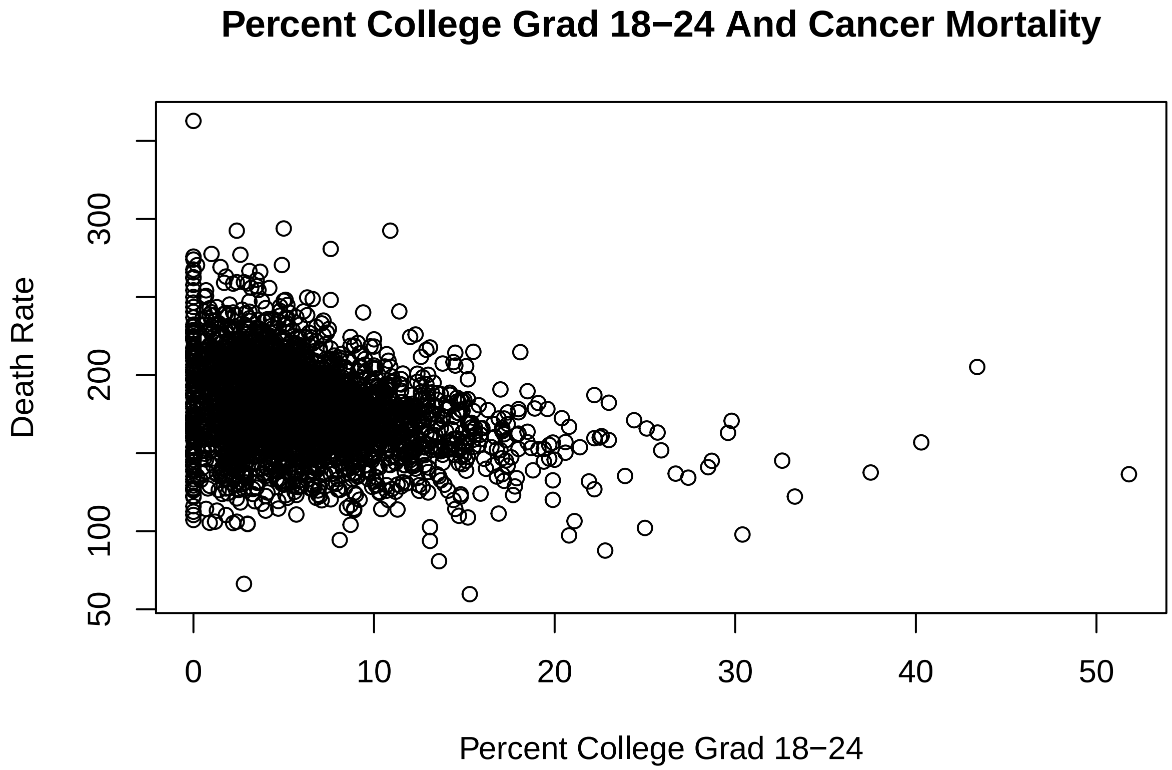


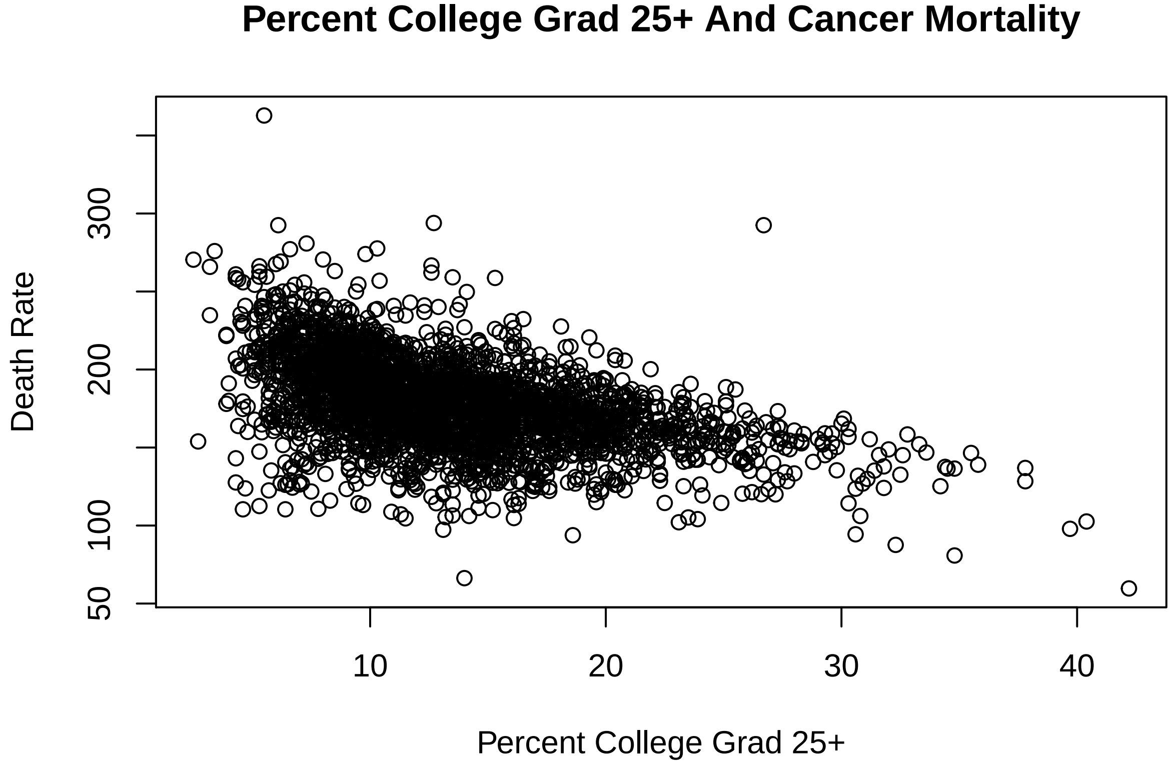
Percent HS Education 25+ And Cancer Mortality



Percent Some College 18−24 And Cancer Mortality

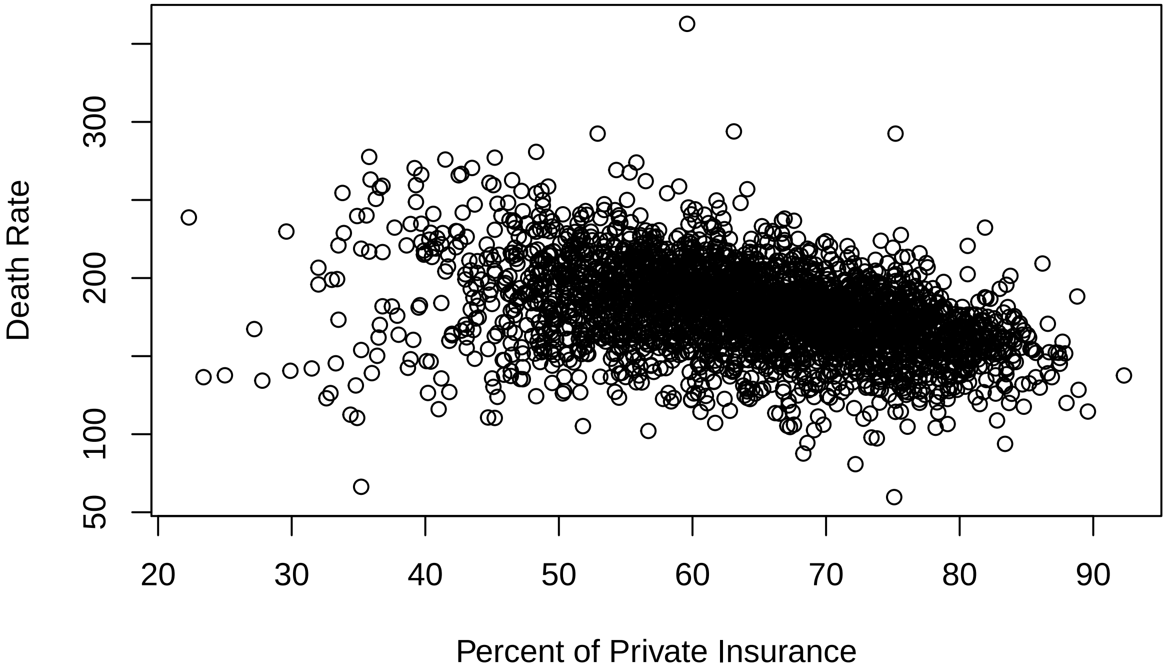




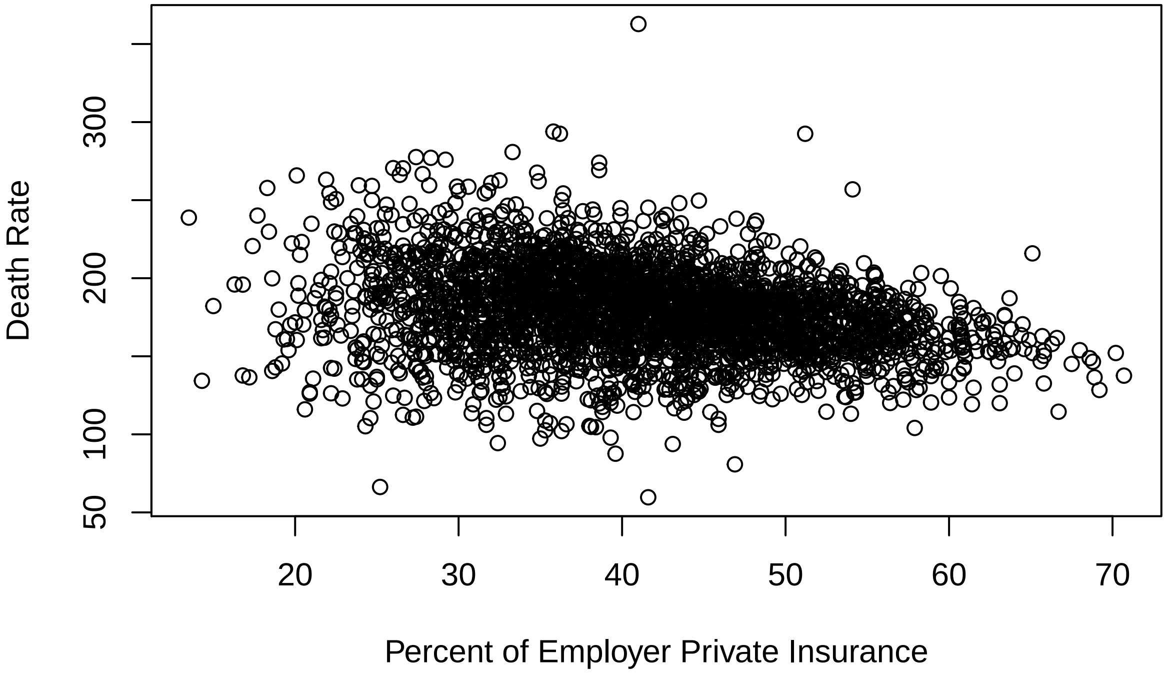


Conclusion of Education Variables: - As a population is more educated, cancer mortality falls - It seems that college grauates make up less percent of cancer moratility population

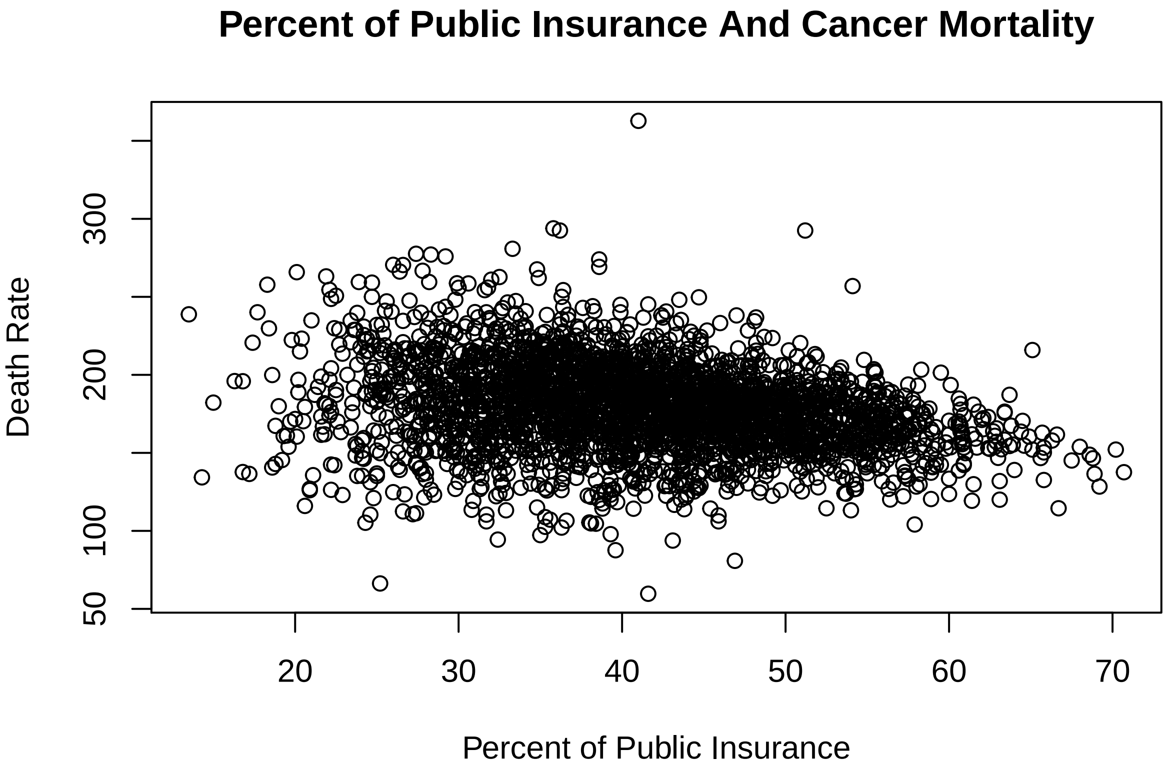
Insurance Coverage: PctPrivateCoverage, PctEmpPrivCoverage, PctPublicCoverage



Percent of Private Insurance And Cancer Mortality



Percent of Employer Private Insurance And Cancer Mortality



Conclusion on Insurance Variables: - It seems when a population has private insurance cancer mortality is down - When a population has public insurance, cancer mortality is up - Public coverage could be correlated by income/poverty/unemployment