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"
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????"

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.
4)