w203_lab1_Cancer_EDA

Nikita Nadkarni, Thomas Drage, Venkatesh Nagapudi

September 14, 2018

W203 Statistics for Data Science

Lab 1 Cancer EDA

1. Introduction (Just to kick things off)

This is an exploratory data analysis to examine the relationship between patient death rates caused by cancer versus various factors including the type of patient coverage, patient income levels and so on. The data analysis includes the following sections: 1. Preliminary Dataset analysis 2. Problems with the Dataset 3. Conclusions that can be reasonably achieved 4. Recommendations 5. Improvements to the Dataset

First include library for scatterplots

```
library(car)
```

```
## Loading required package: carData
```

```
cancer_data = read.csv("cancer.csv")
```

List of variables in cancer data

```
(list_of_varibles = objects(cancer_data))
```

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

A high level summary of cancer_data shows 3047 observations of 30 variables. The important dependent variable is the deathRate. A close second is the incidence rate of cancer, which is the "avgAnnCount". Important indepdent variables include how poor a patient is ("povertyPercent"), whether he/she has Private or Public coverage ("PctPrivateCoverage", "PctPublicCoverage"), his/her race and so on. (need to add more)

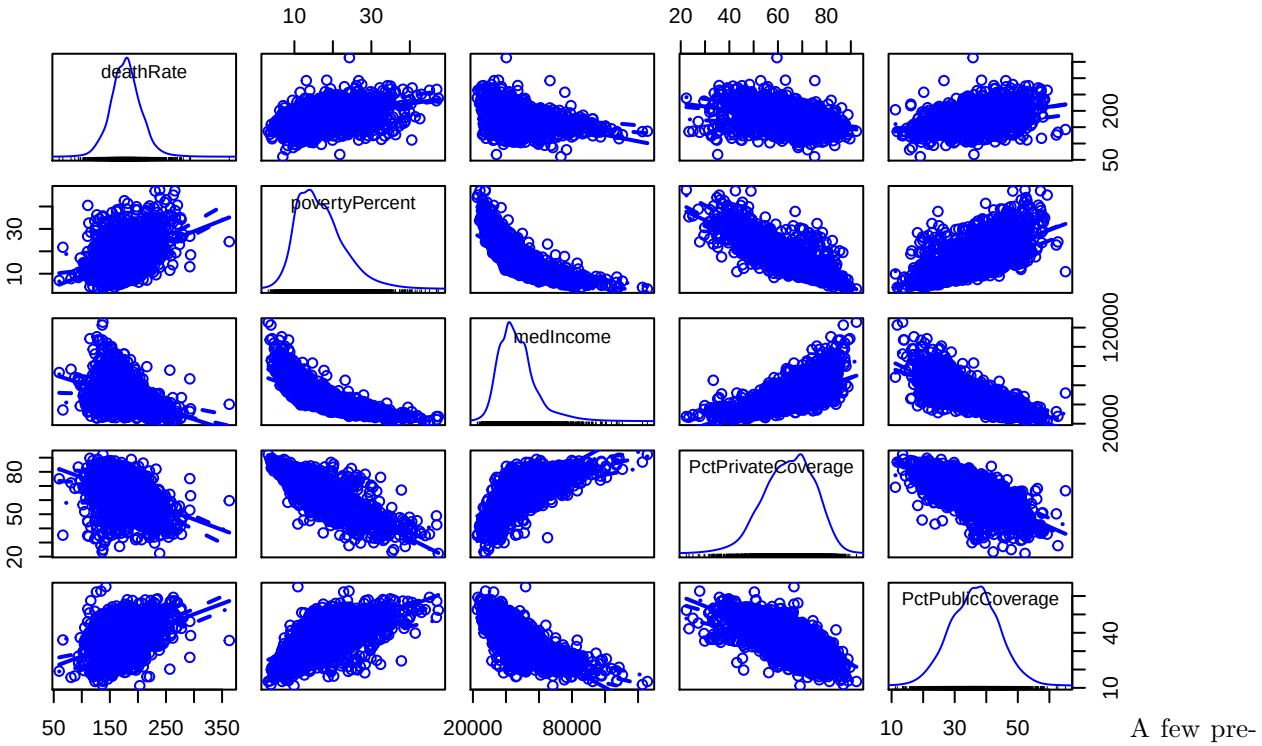
```
str(cancer_data)
```

```
## 'data.frame':    3047 obs. of  30 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ avgAnnCount         : num  1397 173 102 427 57 ...
##  $ medIncome           : int  61898 48127 49348 44243 49955 52313 37782 40189 42579 60397 ...
##  $ popEst2015          : int  260131 43269 21026 75882 10321 61023 41516 20848 13088 843954 ...
##  $ povertyPercent      : num  11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
```

```
##  $ binnedInc           : Factor w/ 10 levels "(34218.1, 37413.8]",..: 9 6 6 4 6 7 2 2 3 8 ...
##  $ MedianAge           : num  39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
##  $ MedianAgeMale       : num  36.9 32.2 44 42.2 47.8 43.5 42.2 50.8 48.4 34.7 ...
##  $ MedianAgeFemale     : num  41.7 33.7 45.8 43.4 48.9 48 43.5 52.5 49.8 37 ...
##  $ Geography           : Factor w/ 3047 levels "Abbeville County, South Carolina",..: 1459 1460 1464
##  $ AvgHouseholdSize    : num  2.54 2.34 2.62 2.52 2.34 2.58 2.42 2.24 2.38 2.65 ...
##  $ PercentMarried      : num  52.5 44.5 54.2 52.7 57.8 50.4 54.1 52.7 55.9 50 ...
##  $ PctNoHS18_24        : num  11.5 6.1 24 20.2 14.9 29.9 26.1 27.3 34.7 15.6 ...
##  $ PctHS18_24          : num  39.5 22.4 36.6 41.2 43 35.1 41.4 33.9 39.4 36.3 ...
##  $ PctSomeCol18_24     : num  42.1 64 NA 36.1 40 NA NA 36.5 NA NA ...
##  $ PctBachDeg18_24     : num  6.9 7.5 9.5 2.5 2 4.5 5.8 2.2 1.4 7.1 ...
##  $ PctHS25_Over        : num  23.2 26 29 31.6 33.4 30.4 29.8 31.6 32.2 28.8 ...
##  $ PctBachDeg25_Over   : num  19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
##  $ PctEmployed16_Over  : num  51.9 55.9 45.9 48.3 48.2 44.1 51.8 40.9 39.5 56.6 ...
##  $ PctUnemployed16_Over: num  8 7.8 7 12.1 4.8 12.9 8.9 8.9 10.3 9.2 ...
##  $ PctPrivateCoverage  : num  75.1 70.2 63.7 58.4 61.6 60 49.5 55.8 55.5 69.9 ...
##  $ PctEmpPrivCoverage  : num  41.6 43.6 34.9 35 35.1 32.6 28.3 25.9 29.9 44.4 ...
##  $ PctPublicCoverage   : num  32.9 31.1 42.1 45.3 44 43.2 46.4 50.9 48.1 31.4 ...
##  $ PctWhite            : num  81.8 89.2 90.9 91.7 94.1 ...
##  $ PctBlack            : num  2.595 0.969 0.74 0.783 0.27 ...
##  $ PctAsian            : num  4.822 2.246 0.466 1.161 0.666 ...
##  $ PctOtherRace        : num  1.843 3.741 2.747 1.363 0.492 ...
##  $ PctMarriedHouseholds: num  52.9 45.4 54.4 51 54 ...
##  $ BirthRate           : num  6.12 4.33 3.73 4.6 6.8 ...
##  $ deathRate           : num  165 161 175 195 144 ...
```

How is death_rate correlated to the important variables? This ScatterPlotMatrix might throw some light

```
scatterplotMatrix(~ deathRate + povertyPercent + medIncome + PctPrivateCoverage + PctPublicCoverage, da
```



A few preliminary observations show that deathRate is positively correlated to povertyPercent and PctPublicCoverage, while it is negatively correlated to medIncome and PctPrivateCoverage. This can be verified with the

correlations below:

```
#correlation of deathRate to important variables
cor(cancer_data$deathRate, cancer_data$povertyPercent)
```

```
## [1] 0.429389
```

```
cor(cancer_data$deathRate, cancer_data$medIncome)
```

```
## [1] -0.4286149
```

```
cor(cancer_data$deathRate, cancer_data$PctPrivateCoverage)
```

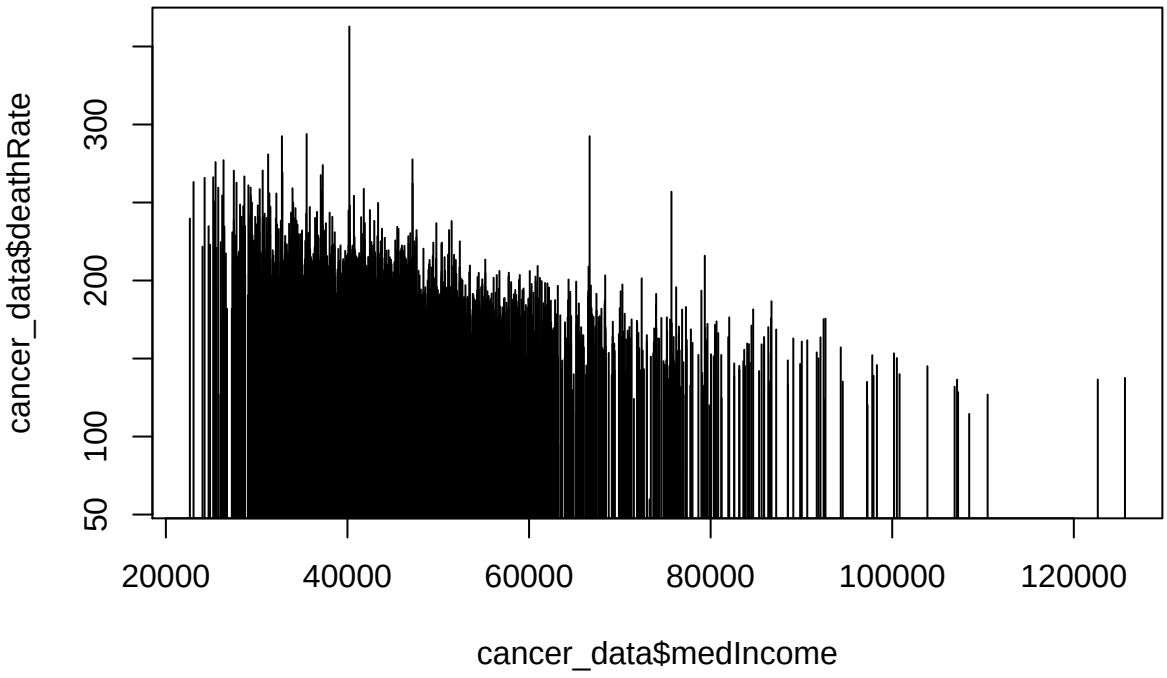
```
## [1] -0.3860655
```

```
cor(cancer_data$deathRate, cancer_data$PctPublicCoverage)
```

```
## [1] 0.4045717
```

There seems to be a high degree of correlation between income and deathRate. Is it possible that the deathRates are high because the cancer can be suppressable through treatment and the poorer people cannot afford it? Or is it because the cancer is incurable? We are not sure that this distinction can be reasonably figured out. But let's say that the cancer is curable - then there should be a correlation between deathRate and income. Let's see:

```
plot (~ cancer_data$medIncome + cancer_data$deathRate, type = "h")
```



There seems to be a gradual drop in death rates for people with higher median incomes. However, there are definitely some outliers as we see a lot of spikes in the death rates in some cases. Are these related to incurable cancers? Are these related to the geography in some ways? How do we figure out?

2. Problems with the Dataset

Thomas/Nikita - am using this section to add any problems with the dataset. We have to figure out a coherent way to structure this analysis and I am fine with any way you suggest.

a) How do we determine whether the deaths were caused because the cancer was undiagnosed? Or whether it was diagnosed and not cured because it was incurable? Or whether it was diagnosed and not cured because the patient couldn't afford it? Since everything is averaged out per county, there is no way of knowing. Perhaps in my mind, the biggest issue with this dataset is that it provides the summary data per geography, rather than the individual patient data itself. What we need is per patient data, not the summary data... perhaps we need to bring this up as the major problem. Thoughts?

b) PctPrivateCoverage and PctPublicCoverage don't add up to 100%. So some people are covered both with private and public coverage or some don't have coverage at all. The numbers range from 65 to 131 when you add them up.

```
head(cancer_data$PctPublicCoverage + cancer_data$PctPrivateCoverage, 25)
```

```
##  [1] 108.0 101.3 105.8 103.7 105.6 103.2  95.9 106.7 103.6 101.3 108.8
## [12] 102.5 100.5  99.2 101.7 104.1 105.2 114.4 103.7 100.6 101.2  90.3
## [23] 100.4 103.1 106.5
```

```
min(cancer_data$PctPublicCoverage + cancer_data$PctPrivateCoverage)
```

```
## [1] 65.4
```

```
max(cancer_data$PctPublicCoverage + cancer_data$PctPrivateCoverage)
```

```
## [1] 131.7
```

c) The same problem exists with the race related information. It doesn't add up to 100%.

```
head(cancer_data$PctAsian + cancer_data$PctBlack + cancer_data$PctWhite + cancer_data$PctOtherRace, 25)
```

```
##  [1] 91.04059 96.18520 94.87512 95.05131 95.53218 91.38853 84.94629
##  [8] 93.88714 92.79766 90.18118 96.89145 94.51078 96.01530 93.19282
## [15] 94.21482 91.13622 92.48829 94.22553 95.14114 92.70208 94.96758
## [22] 92.90322 98.08738 97.09264 99.20833
```

```
min(cancer_data$PctAsian + cancer_data$PctBlack + cancer_data$PctWhite + cancer_data$PctOtherRace)
```

```
## [1] 11.22511
```

```
max(cancer_data$PctAsian + cancer_data$PctBlack + cancer_data$PctWhite + cancer_data$PctOtherRace)
```

```
## [1] 100
```

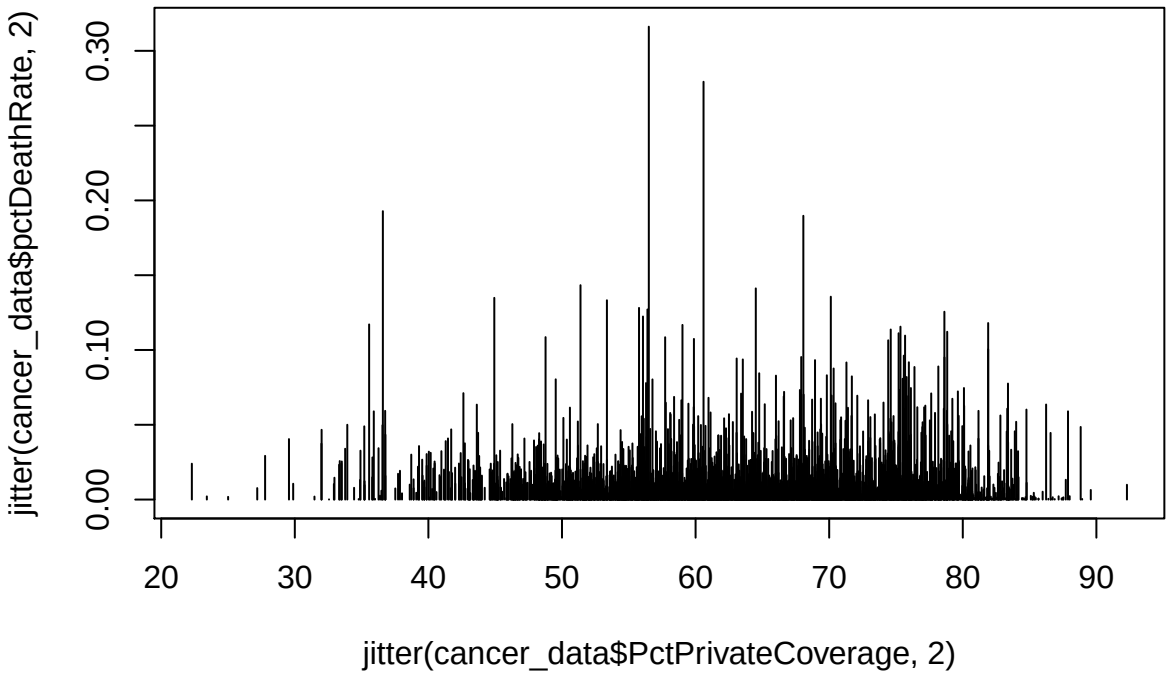
d) deathRate doesn't take into account the population of the county. Perhaps this is a huge problem. Maybe the data is not so bad if you find the actual % of population that died because of cancer and correlate it against other variables? A lot of the data so far that I gathered might be wrong if the main variable is not meaningful. Thoughts?

```
#added new column into cancer_data
cancer_data$pctDeathRate = cancer_data$deathRate/cancer_data$popEst2015
```

Also, how do we know that the deathRate data is for 2015?

Let's see how pctDeathRate vs PcctPrivateCoverage looks now

```
plot (~ jitter(cancer_data$PctPrivateCoverage,2) + jitter(cancer_data$pctDeathRate,2), type = "h")
```



How about the correlation between the two variables?

```
cor(cancer_data$pctDeathRate, cancer_data$PctPrivateCoverage)
```

```
## [1] -0.06555567
```

This shows very little correlation. (Nikita/Thomas - thoughts??) How about the correlation between pctDeathRates and PublicCoverage?

```
cor(cancer_data$pctDeathRate, cancer_data$PctPublicCoverage)
```

```
## [1] 0.1291706
```

This seems somewhat intuitive. Perhaps with public coverage, treatment is bad.

How about median income?

```
cor(cancer_data$pctDeathRate, cancer_data$medIncome)
```

```
## [1] -0.167507
```

A negative correlation that is as strong as the public coverage. Perhaps makes sense.

And how about poverty level

```
cor(cancer_data$pctDeathRate, cancer_data$povertyPercent)
```

```
## [1] 0.04147443
```

Not as correlated... what does this mean? That deathrates are not dependent so much on poverty?