Missing Values Alex

Alex

08/02/2022

```
load("cancer.rdata")
summary(cancer)
```

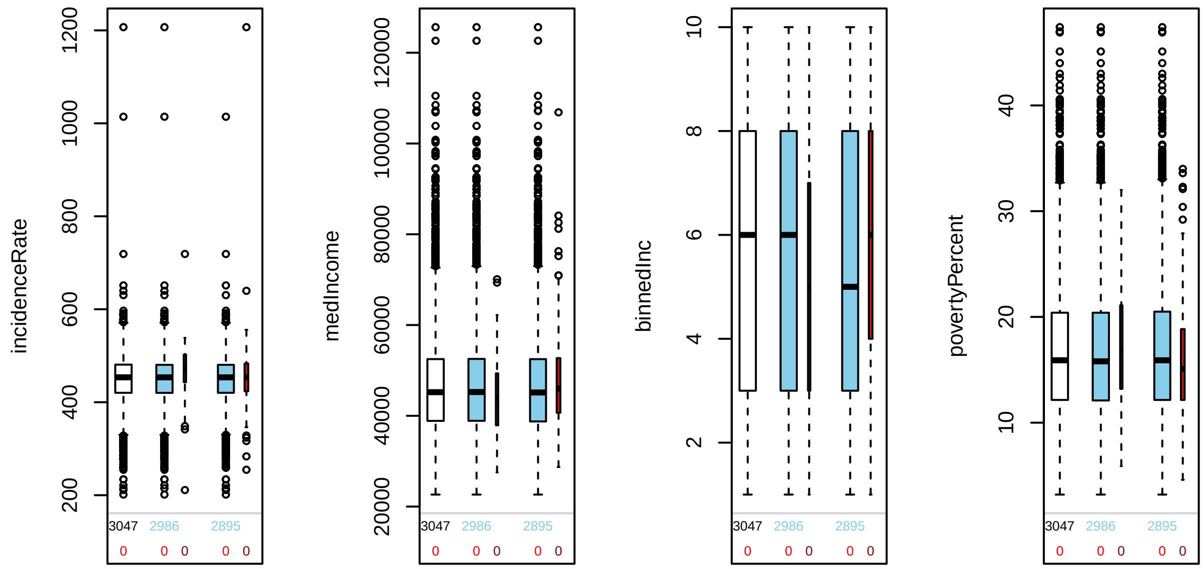
```
##   Geography          incidenceRate      medIncome                   binnedInc
##  Length:3047        Min.   : 201.3   Min.   : 22640   [22640, 34218.1]  : 306
##  Class :character   1st Qu.: 420.3   1st Qu.: 38883   (45201, 48021.6]  : 306
##  Mode  :character   Median : 453.5   Median : 45207   (54545.6, 61494.5]: 306
##                     Mean   : 448.3   Mean   : 47063   (42724.4, 45201]  : 305
##                     3rd Qu.: 480.9   3rd Qu.: 52492   (48021.6, 51046.4]: 305
##                     Max.   :1206.9   Max.   :125635   (51046.4, 54545.6]: 305
##                                                       (Other)           :1214
##  povertyPercent  MedianAgeMale   MedianAgeFemale AvgHouseholdSize
##  Min.   : 3.20   Min.   :22.40   Min.   :22.30   Min.   :0.0221
##  1st Qu.:12.15   1st Qu.:36.35   1st Qu.:39.10   1st Qu.:2.3700
##  Median :15.90   Median :39.60   Median :42.40   Median :2.5000
##  Mean   :16.88   Mean   :39.57   Mean   :42.15   Mean   :2.4797
##  3rd Qu.:20.40   3rd Qu.:42.50   3rd Qu.:45.30   3rd Qu.:2.6300
##  Max.   :47.40   Max.   :64.70   Max.   :65.70   Max.   :3.9700
##
##  PercentMarried  PctEmployed16_Over PctUnemployed16_Over PctPrivateCoverage
##  Min.   :23.10   Min.   :17.60      Min.   : 0.400       Min.   :22.30
##  1st Qu.:47.75   1st Qu.:48.60      1st Qu.: 5.500       1st Qu.:57.20
##  Median :52.40   Median :54.50      Median : 7.600       Median :65.10
##  Mean   :51.77   Mean   :54.15      Mean   : 7.852       Mean   :64.35
##  3rd Qu.:56.40   3rd Qu.:60.30      3rd Qu.: 9.700       3rd Qu.:72.10
##  Max.   :72.50   Max.   :80.10      Max.   :29.400       Max.   :92.30
##                  NA's   :152
##  PctEmpPrivCoverage PctPublicCoverage    PctBlack       PctMarriedHouseholds
##  Min.   :13.5       Min.   :11.20     Min.   : 0.0000   Min.   :22.99
##  1st Qu.:34.5       1st Qu.:30.90     1st Qu.: 0.6207   1st Qu.:47.76
##  Median :41.1       Median :36.30     Median : 2.2476   Median :51.67
##  Mean   :41.2       Mean   :36.25     Mean   : 9.1080   Mean   :51.24
##  3rd Qu.:47.7       3rd Qu.:41.55     3rd Qu.:10.5097   3rd Qu.:55.40
##  Max.   :70.7       Max.   :65.10     Max.   :85.9478   Max.   :78.08
##
##     Edu18_24       deathRate
##  Min.   :1.487   Min.   : 59.7
##  1st Qu.:2.206   1st Qu.:161.2
##  Median :2.340   Median :178.1
##  Mean   :2.347   Mean   :178.7
##  3rd Qu.:2.486   3rd Qu.:195.2
##  Max.   :3.307   Max.   :362.8
##
```

There are some missing values in PctEmployed16_Over which need to be checked. Before that is checked though we should note that the outliers in AvgHouseholdSize should be treated as missing values as they are wrongly inputted (reference).

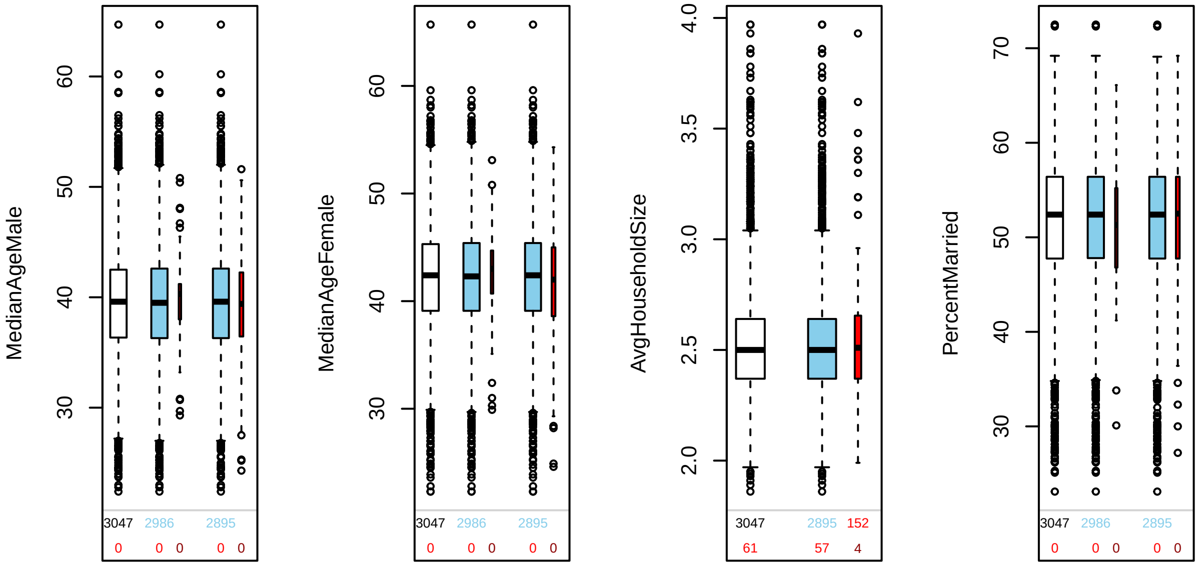
```
cancer1 <- cancer
cancer1$AvgHouseholdSize[cancer1$AvgHouseholdSize < 0.5] <- NA
```

Now I use the VIM package and the pbox() function to show that the missing data are all MCAR and can thus be easily dealt with.

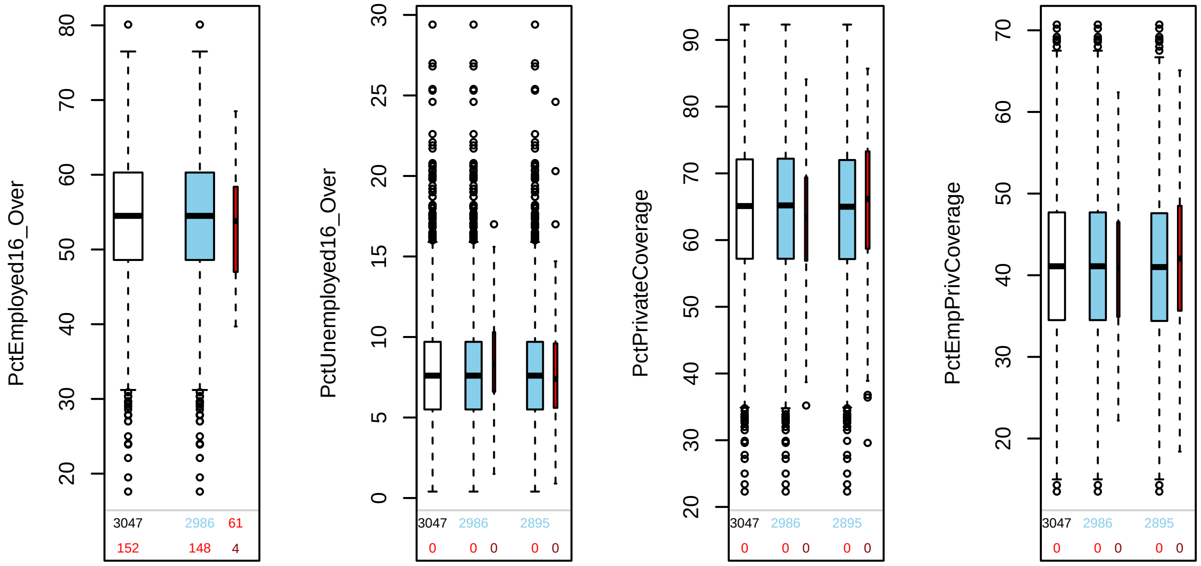
```
library(VIM)
par(mfrow = c(1,4))
for(i in 2:5){
  pbox(cancer1, pos = i)
}
```



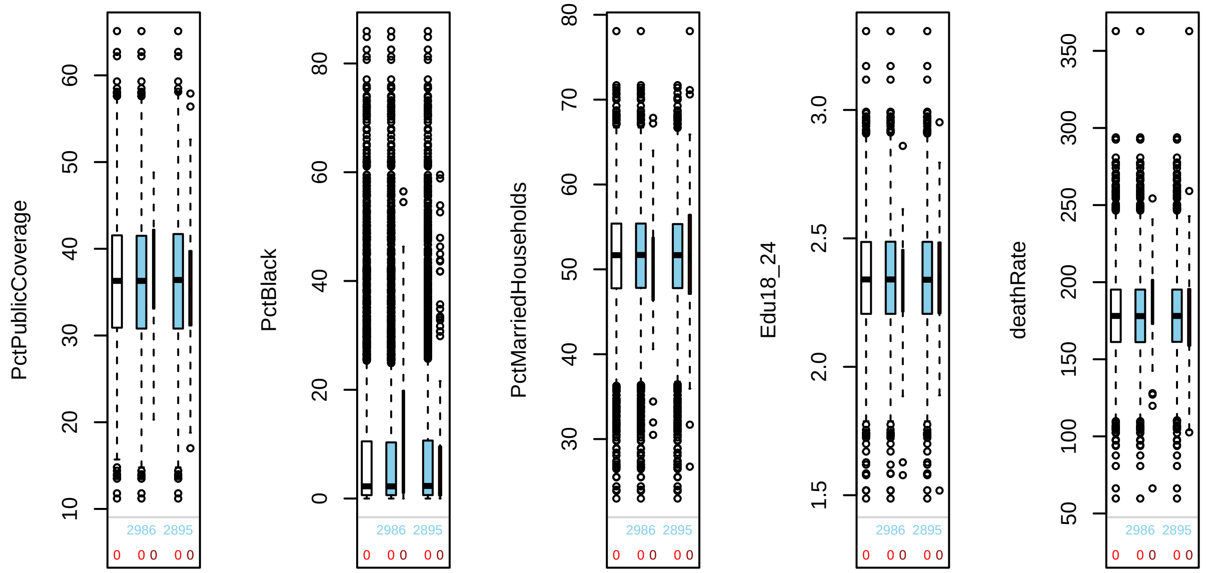
```
par(mfrow = c(1,4))
for(i in 6:9){
  pbox(cancer1, pos = i)
}
```



```r
par(mfrow = c(1,4))
for(i in 10:13){
  pbox(cancer1, pos = i)
}
```



```
par(mfrow = c(1,5))
for(i in 14:18){
  pbox(cancer1, pos = i)
}
```



From the pbox plots we have evidence that the missing values interrogated are MCAR. Since the proportion of rows with missing values is small I recommend simply deleting the rows with missing values as it should not affect the validity of our analysis on the whole dataset as the smaller sample is still representative.

```
cancer2 <- na.omit(cancer1)
```

Our other option is to use the reference to fill in all our other missing values. Despite this being possible it is also un-necessary as the data is MCAR and won't significantly affect our analysis.