ST404 Assignment 1 Alex

ST404 Assignment 1 Alex

Checking the summary and initial EDA

```
##   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.

The minimum value in AvgHouseholdSize is very small which is suspicious and should be immediately investigated.

From the above plot we note that there are many extremely suspicious points with small AvgHouseholdSize.

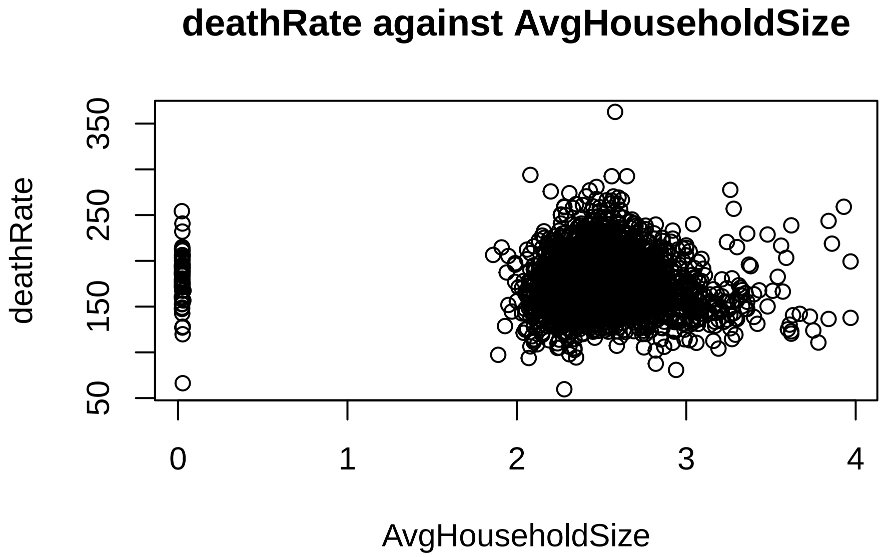


Figure 1: deathRate vs AvgHouseholdSize

We identify one of these points and investigate it:

Geography	AvgHouseholdSize
Berkeley County, West Virginia	0.0263

To check the validity of this data point we find an alternate source of the data at:

https://data.census.gov/cedsci/table?q=average%20household%20size&g=0500000US54003&y=2013&tid=ACSST1Y2013.S1101

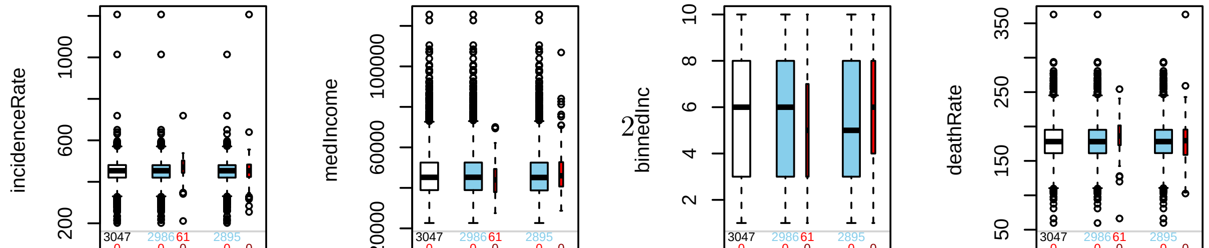
We note that this data recording AvgHouseholdSize in the same year as our data lists the size at 2.61. This is completely different and this is similar for other small values in our dataset.

Hence, these are very likely incorrectly inputted data points and as there is only a small proportion of them we should treat them as missing data and then test to see whether they are MCAR.

```
cancer1 <- cancer
cancer1$AvgHouseholdSize[which(cancer1$AvgHouseholdSize < 0.5)] <- NA
```

Missing values check

Now that we have replaced the small values with NAs we can test the data to see what kind of missing values we have.



With our data we could replace all the data with an alternate source but as the proportion of missing data points is so small and it is MCAR it is safe to just remove the rows with missing data from our data set.

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

My allocation

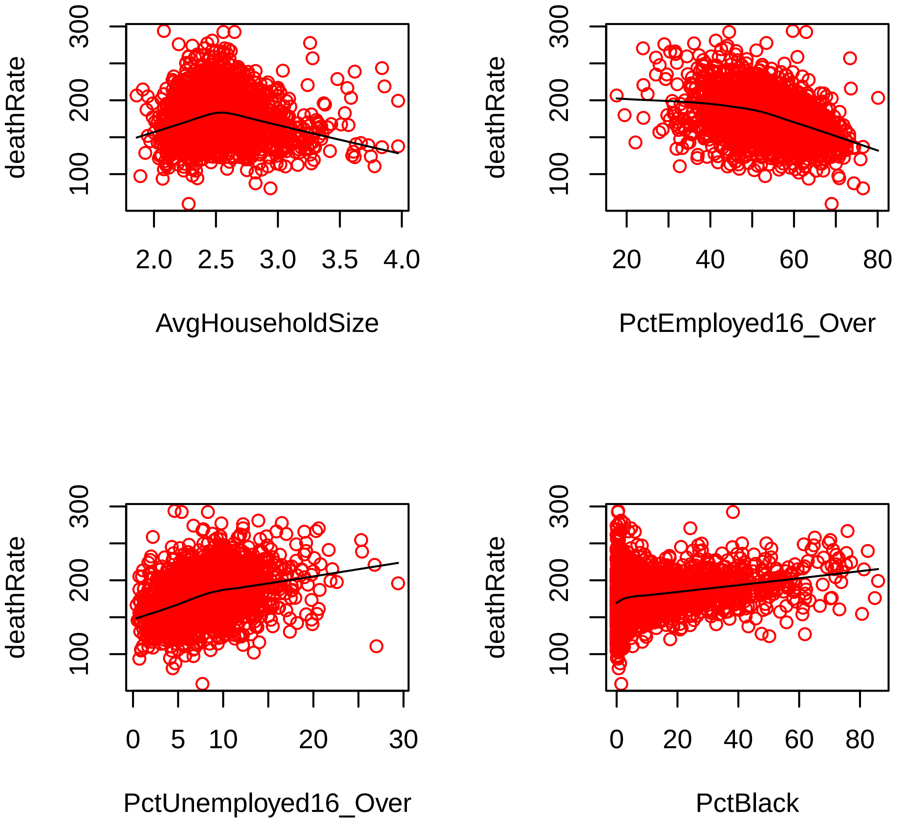


Figure 3: Plots showing deathRate against other variables

From the bivariate plots there is definite heteroskedasticity in pctBlack and for AvgHouseholdSize we see some non linearity. We see a concave shape so advising a more compex model, perhaps with a quadratic term might be advisable as the data is not monotonic.

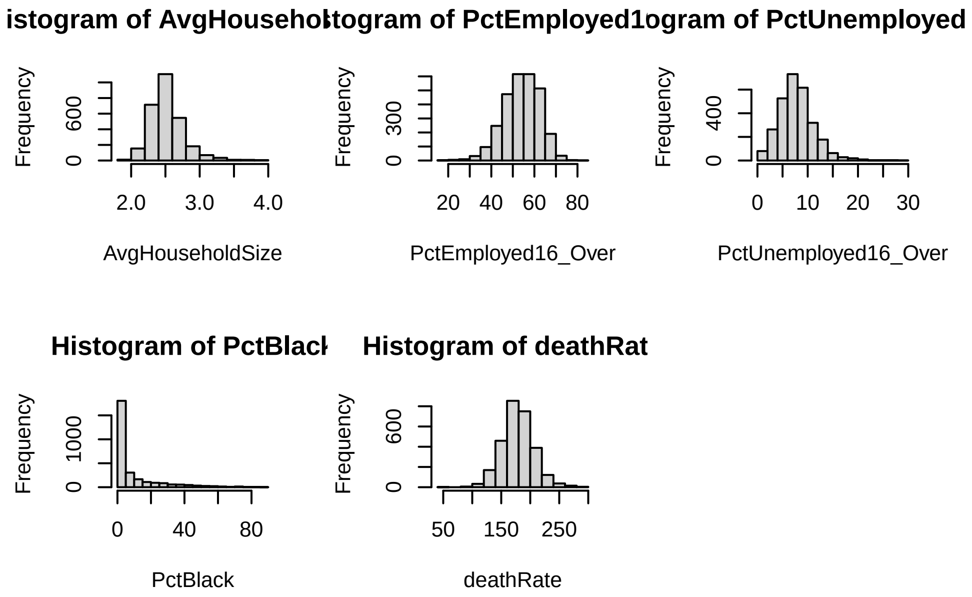


Figure 4: Histograms of our predictor variables

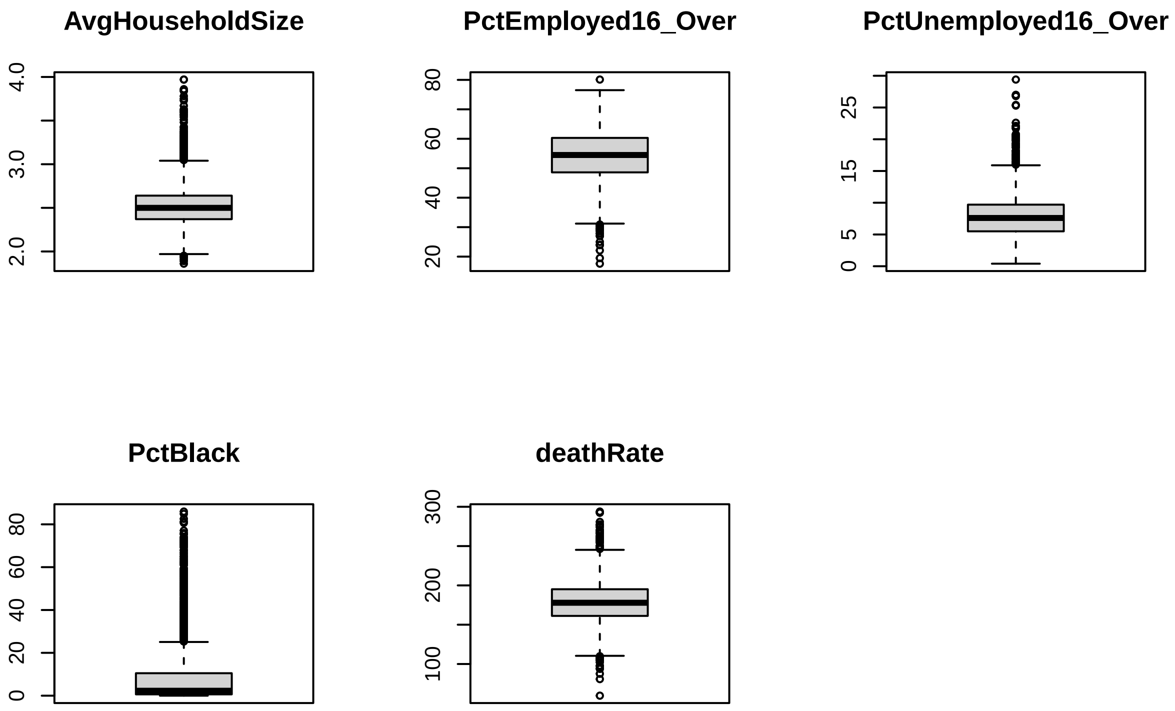


Figure 5: BoxPlots of our variables

Analysis of the above plots

Scatter Plots

From the bivariate plots there is definite heteroskedasticity in pctBlack and for AvgHouseholdSize we see some non linearity. We see a concave shape so advising a more complex model, perhaps with a quadratic term might be advisable as the data is not monotonic.

For heteroskedasticity we would need to perform further tests after fititng a model to check what kind of transformation we'd need to fix it.

From the scatter plots there are no clear outliers, we'd need either some box plots or to look at cook's distance to identify that.

Histograms

Massive right skew for pctBlack. PctUnemployed and AvgHouseholdSize are also a little right skew. I Recommend a log transform for pctBlack and sqrt transforms for pct unemployed and avg household size.

```r
par(mfrow = c(1,3))
with(cancer2, hist(sqrt(AvgHouseholdSize), main = "Transformed AvgHouseholdSize"))
with(cancer2, hist(sqrt(PctUnemployed16_Over), main = "Transformed PctUnemployed16_Over"))
with(cancer2, hist(log(cancer2$PctBlack), main = "Transformed PctBlack"))
```

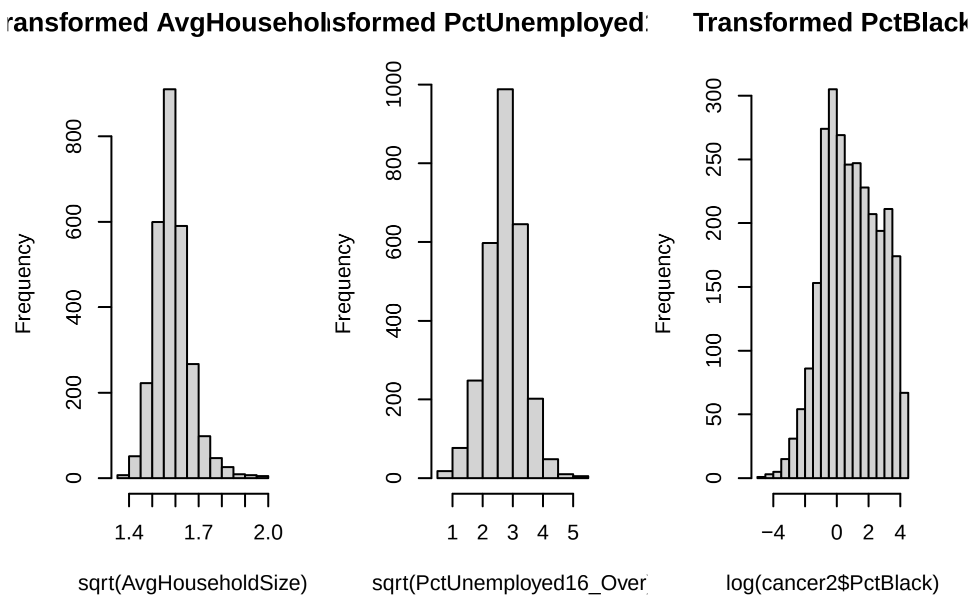


Figure 6: Our transformed histograms

Box Plots

Our Box Plots show we have quite a number of what we would consider outliers accross all our variables. This does not necessarily mean that they should be removed as we do not know their influence yet due to not fitting a model.

We have a severe amount of outliers in PctBlack according to our box plot. This could be due to the very long tail as shown in the scatter plot above.

BIG MAP

From the map we note that the deathRate appears to be higher in the mid-eastern United States