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.

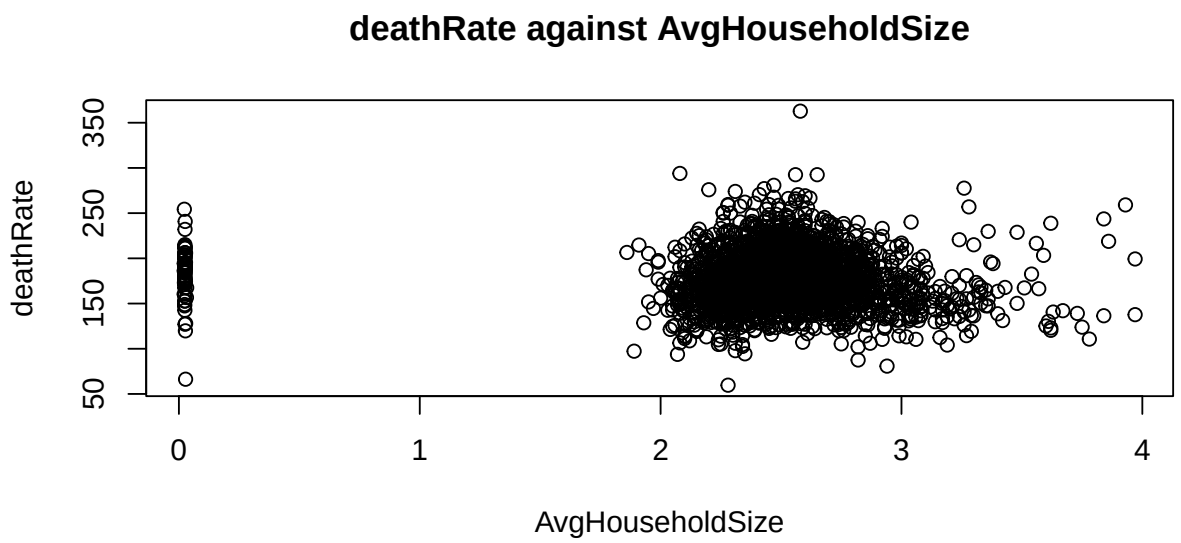


Figure 1: deathRate vs AvgHouseholdSize

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:

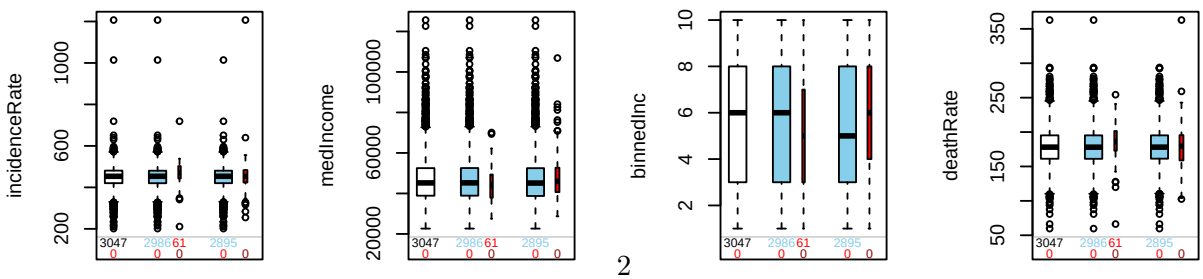
https://data.census.gov/cedsci/table?q=average%20household%20size&g=0500000US54003&y=2013&tid=ACSST1Y2013.S1101

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.

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.



We can see initially that medIncome is right skewed, also median age Male and Female are highly correlated. We will use a pearsons coefficient to check if multicollinearity exist (http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r). binnedInc does not tell us much from this graph - we will instead look at binnedInc and medIncome together as they measure similar data to check for multicollinearity.

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.

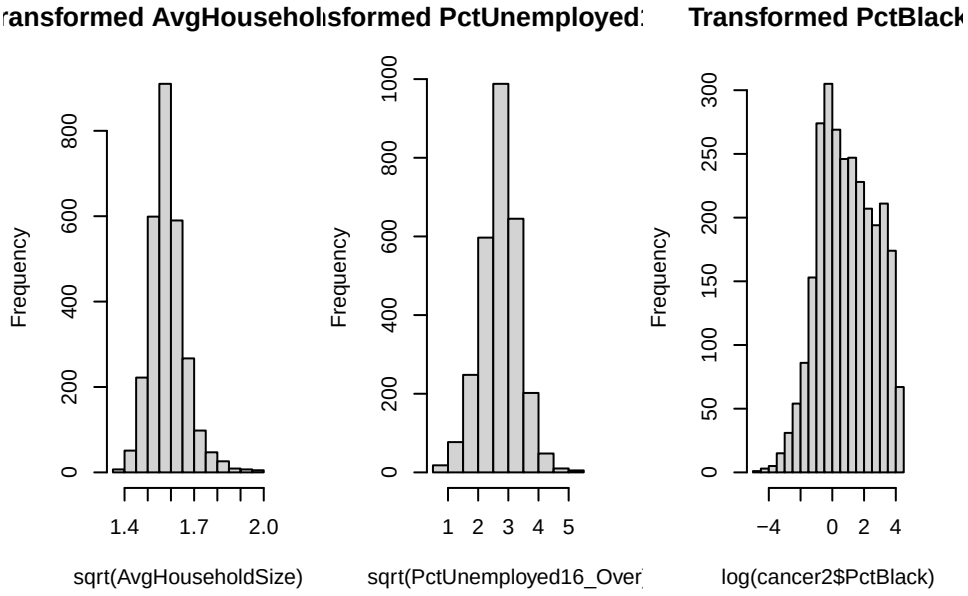
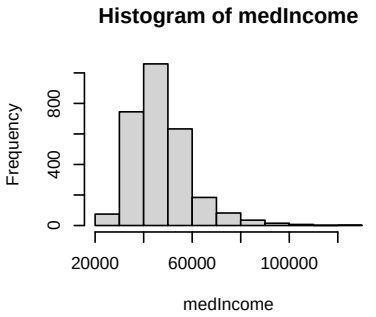
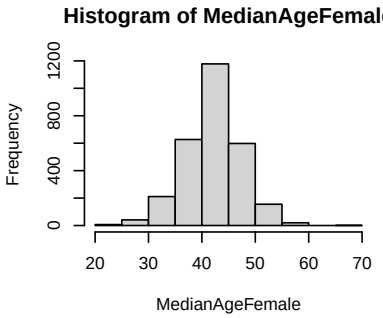
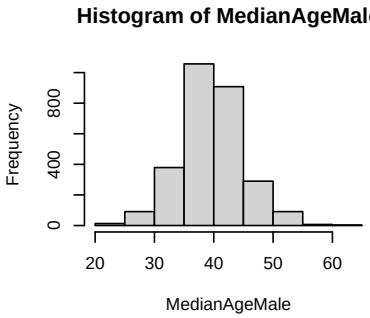
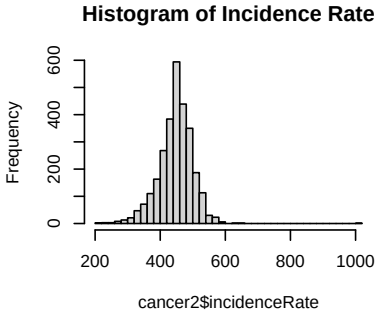


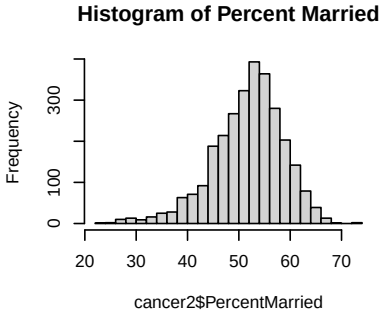
Figure 3: Our transformed histograms

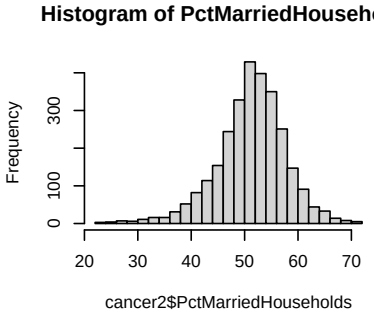

Histogram of medIncome
Frequency
0
400
800
20000
60000
100000
medIncome

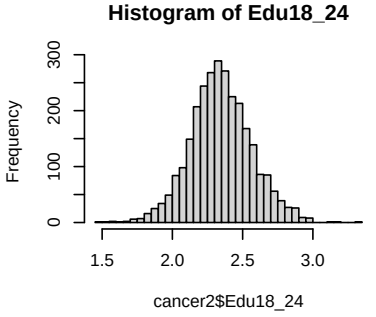

Histogram of MedianAgeFemale
Frequency
0
400
800
1200
20
30
40
50
60
70
MedianAgeFemale

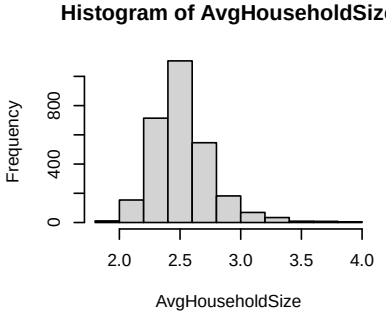

Histogram of MedianAgeMale
Frequency
0
400
800
20
30
40
50
60
MedianAgeMale

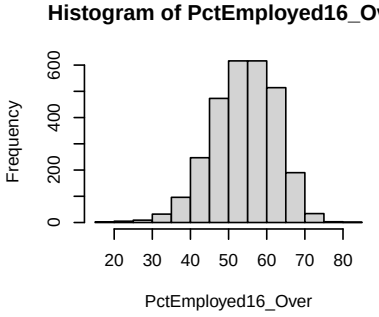

Histogram of Incidence Rate
Frequency
0
200
400
600
200
400
600
800
1000
cancer2$incidenceRate

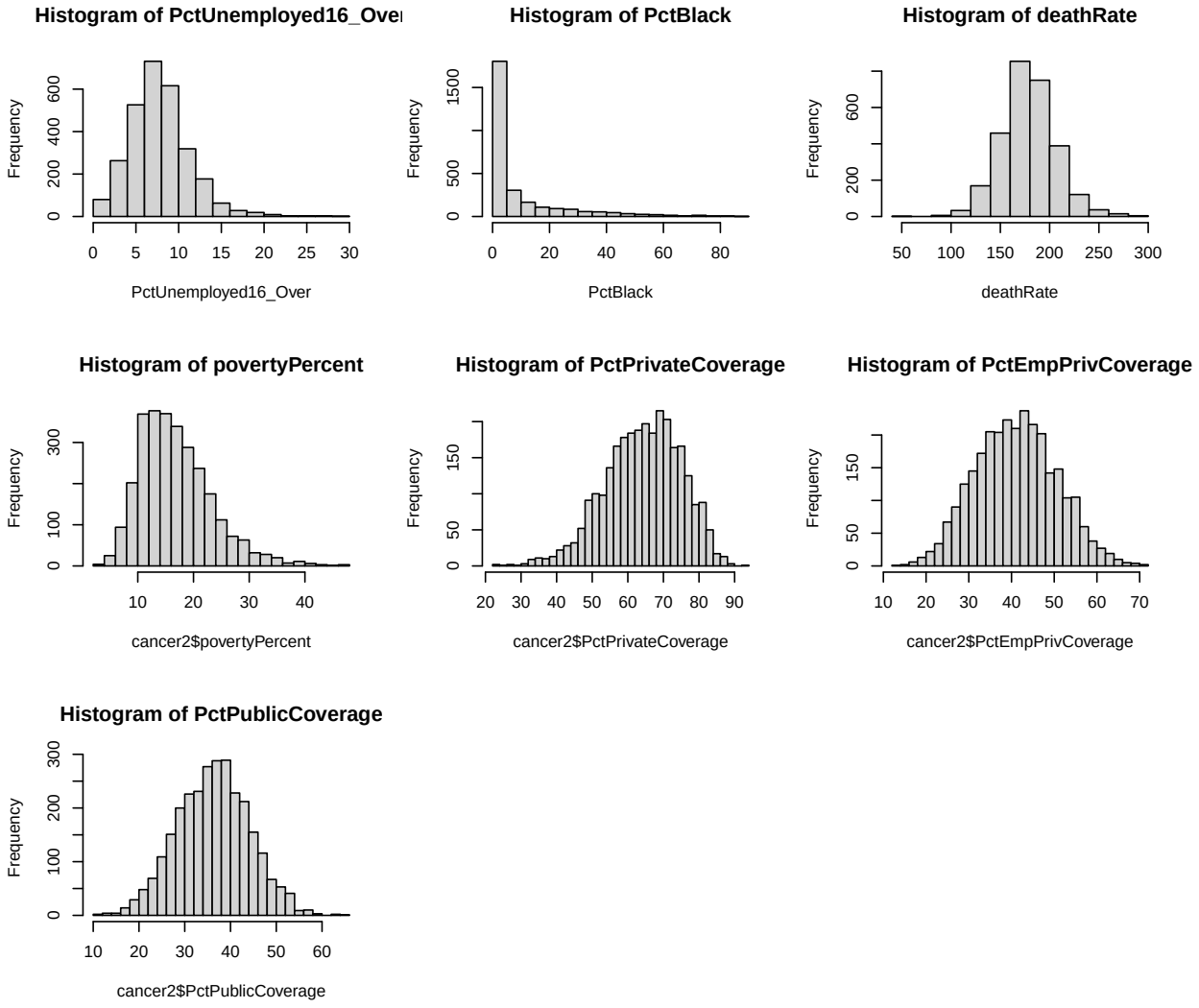

Histogram of Percent Married
Frequency
0
100
300
20
30
40
50
60
70
cancer2$PercentMarried


Histogram of PctMarriedHouseholds
Frequency
0
100
300
20
30
40
50
60
70
cancer2$PctMarriedHouseholds


Histogram of Edu18_24
Frequency
0
100
200
300
1.5
2.0
2.5
3.0
cancer2$Edu18_24

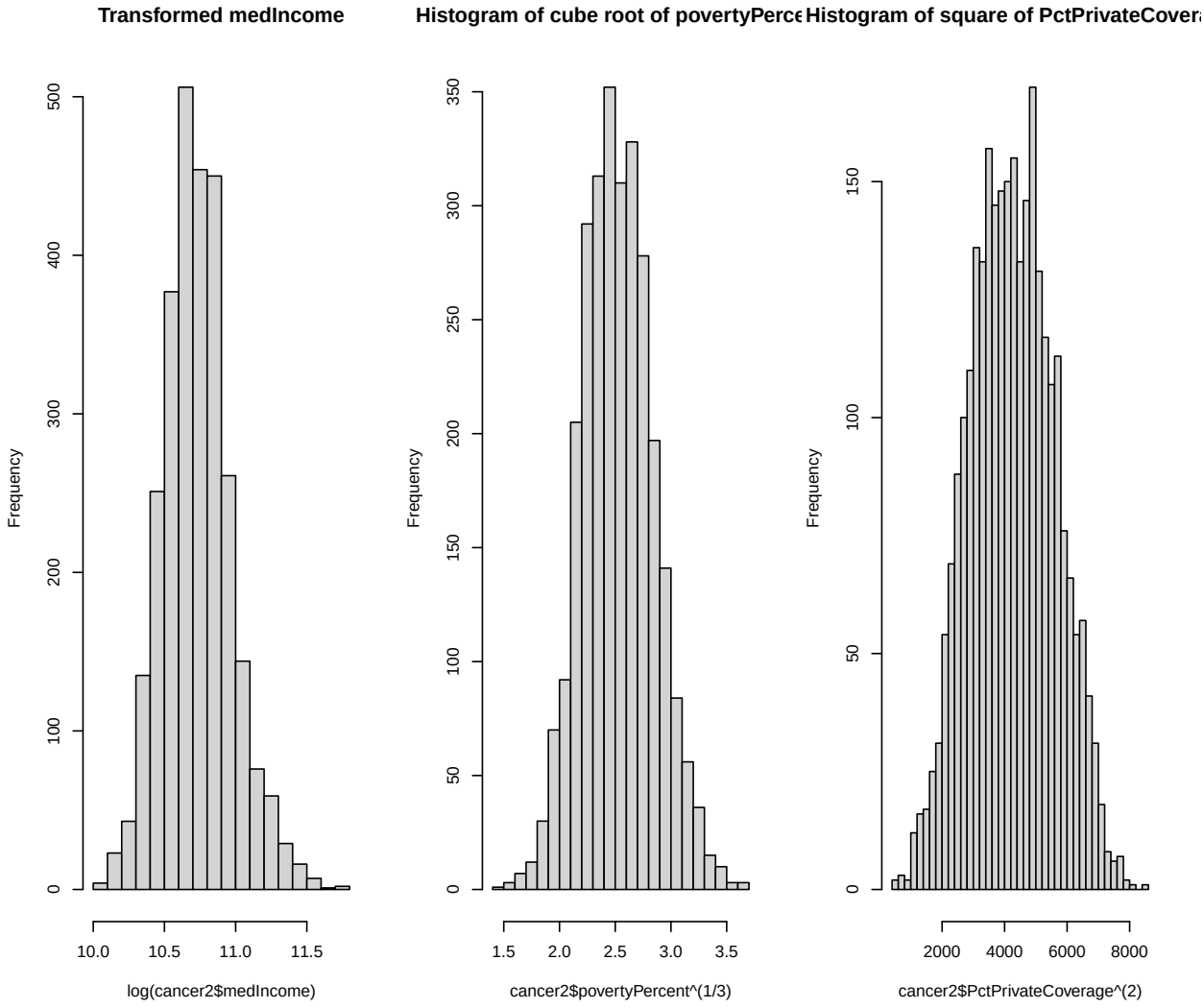

Histogram of AvgHouseholdSize
Frequency
0
400
800
2.0
2.5
3.0
3.5
4.0
AvgHouseholdSize


Histogram of PctEmployed16_Over
Frequency
0
200
400
600
20
30
40
50
60
70
80
PctEmployed16_Over



Massive right skew for medIncome. Median age Male and Female seem to have a pretty symettrical skew. I recommend a log transform for medIncome.

From the histograms, PctPrivateCoverage is slightly left skewed and povertyPercent is right skewed. Transformations are needed. Try cube root for povertyPercent.



There still exists a slight right skew for medIncome but it is an improvement.

The skewness in povertyPercent is removed in the cubeth rooting transfomation.

The skewness is in PctPrivateCoverage is removed by the square transformation.

Bivariate Plots

Plots Against Death Rate

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.

There is definite heteroscedasticity in medIncome and for both MedianAgeFemale and MedianAge-Male 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. We could also consider

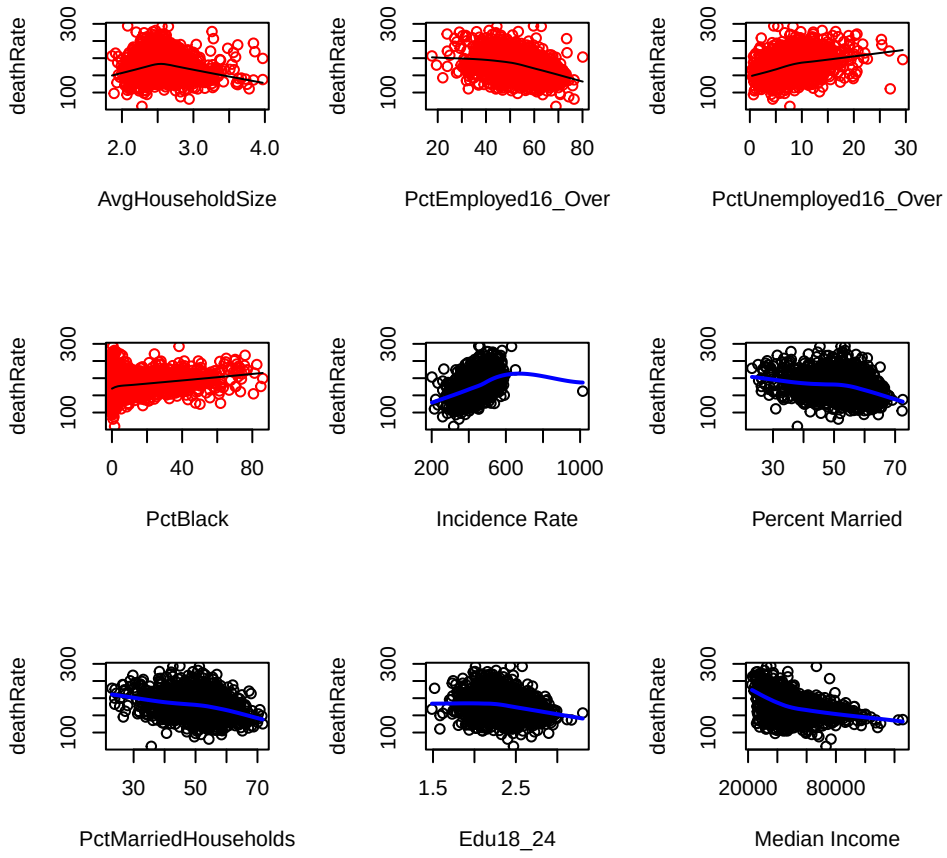


Figure 4: Plots showing deathRate against other variables

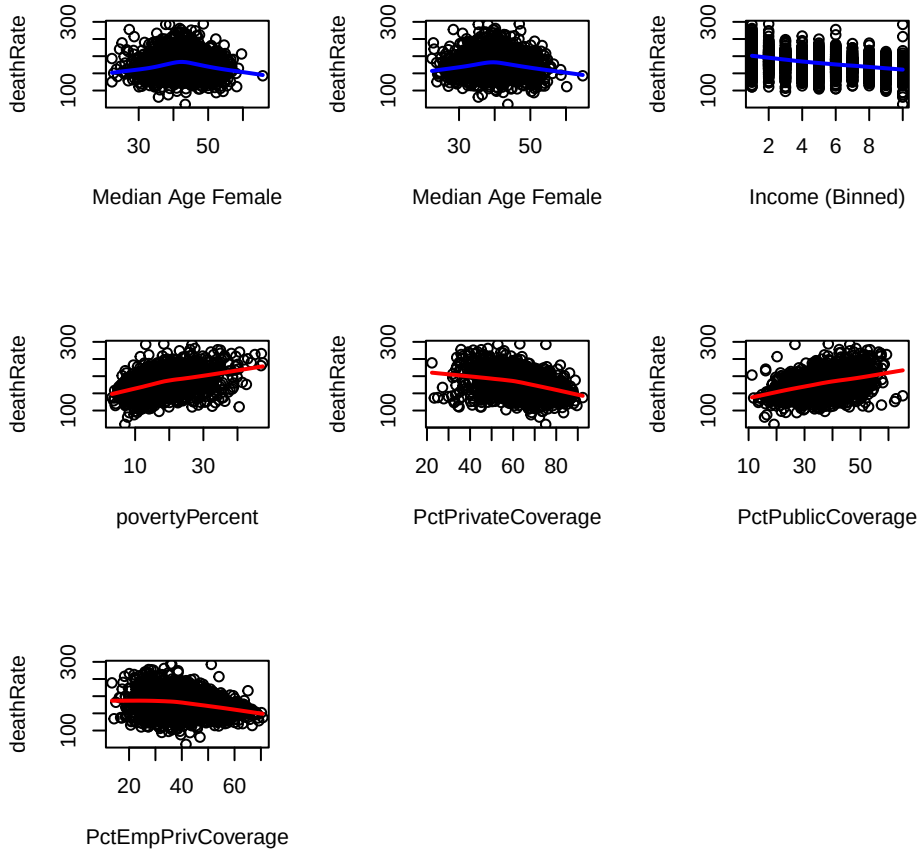


Figure 5: Plots showing deathRate against other variables

combining the two variables by taking an average as their relationship with Death Rate are very similar. BinnedInc does not show us much other than it is linear.

From the scatter plots there are no clear outliers.

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.

From the scatterplots, the predictor variables show a fairly linear relationship with death Rate. The fitted lines for PercentMarried, PctMarriedHouseholds and Edu18_24 show an inverse relationship with the response variable. However,the fitted line for incidence rate indicates that the outliers might have a high influence on the model and therefore we can observe that there is heteroscedasticity. We need to perform further investigation and tests after fitting a model and we can use spreadLevelPlot() to fix heteroscedasticity.

We can see all four variables have fairlty straight fitted lines. This support them being linear. The data points of PctEmpPrivCoverage are getting closer to the fitted line which shows that the variance is decreasing. We might use spreadLevelPlot() to find appropriate power transformation to fix this problem.

Plots for Outliers

Our box plots show we have quite a number of what we would consider outliers accross all our variables apart from binnedInc, which would be impossible due to the bins intervals. 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 medIncome. This is most likely due to natural causes such as a CEO of a large company or a doctor (Reference needed about high paid jobs).

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.

The boxplot for Incidence rate shows the existence of extreme high values.

Therefore, we might want to investigate into the outliers of the predictor variables and decide how we might want to treat them before fitting the model.

Possible options might include deleting the outliers or imputing them.

This might depend on whether the outlier is due to some error (data entry, sampling, measurement) or whether the outlier is natural.

##Remus Boxplot

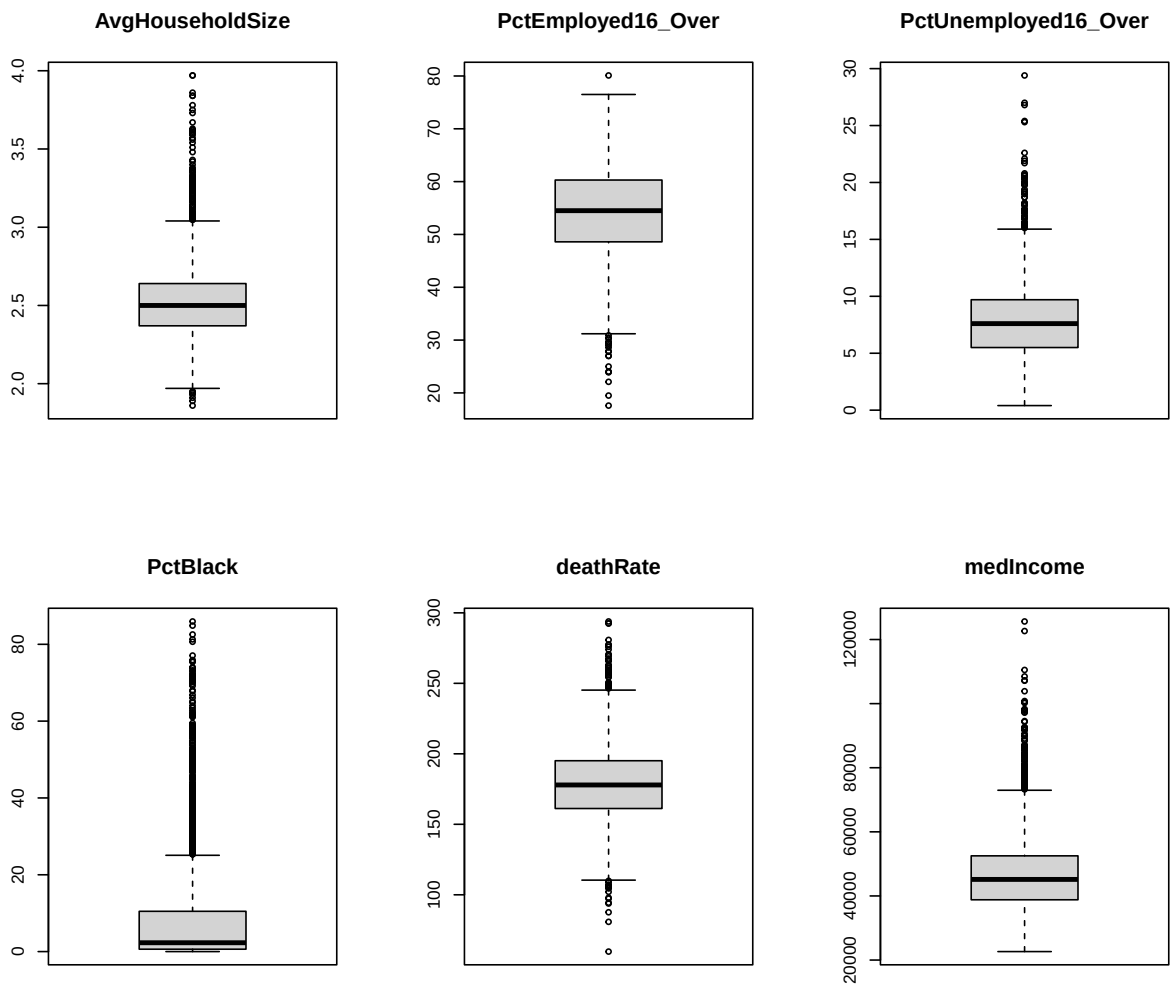


Figure 6: BoxPlots of our variables

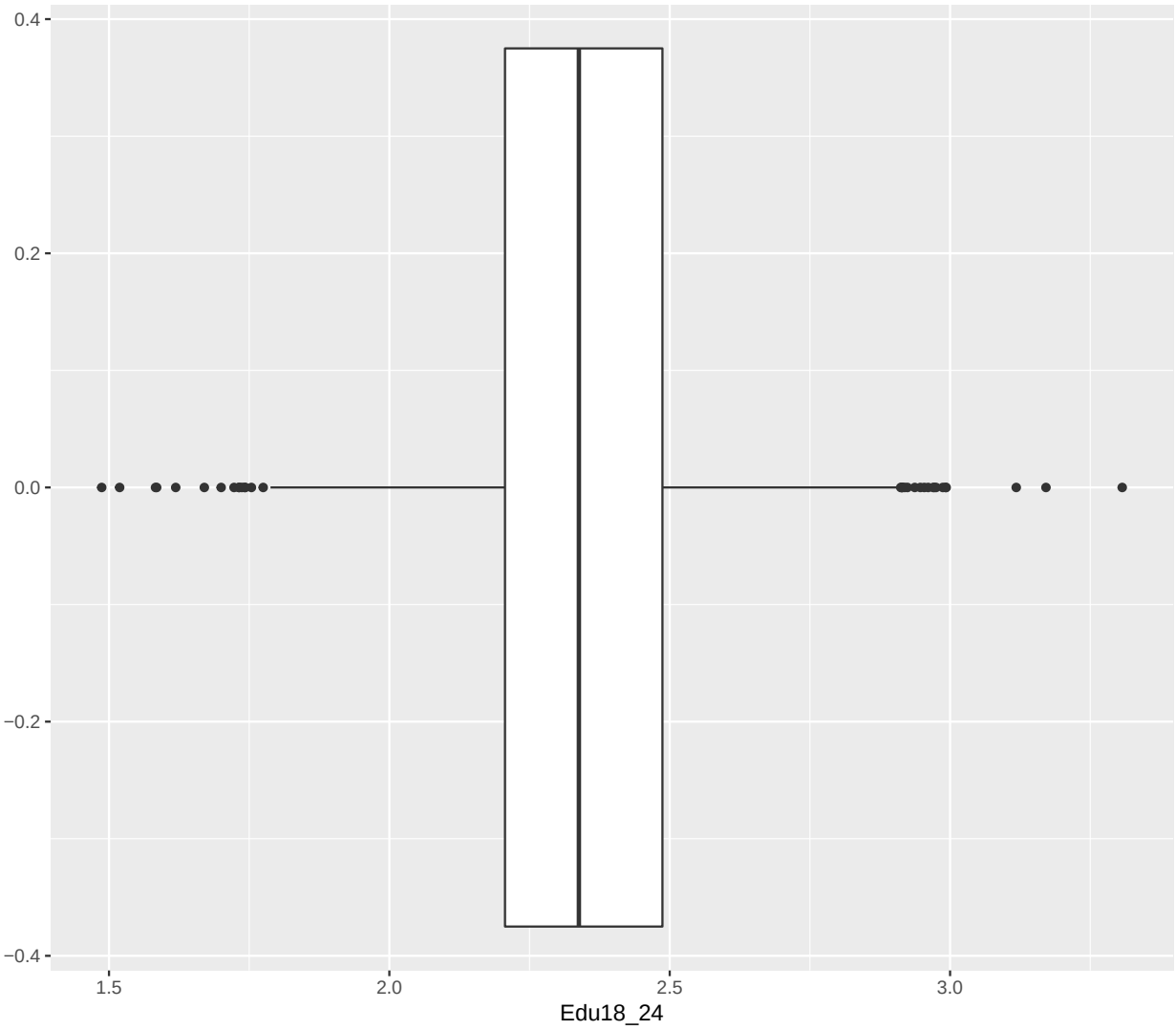


Figure 7: BoxPlots of our variables

All four variables have points lying outside the boxes. Note that PctPrivateCoverage have points lying below the box and povertyPercent have points lying above the box only. This shows potential outliers in particular in PctPrivateCoverage and povertyPercent. Further discoveries and decisions on the outliers should be done when fitting a model.

Multicollinearity

Multicollinearity for Married

```
##             [,1]        [,2]        [,3]        [,4]
## [1,]  1.0000000 -0.11315444 -0.15354148  0.14164656
## [2,] -0.1131544  1.00000000  0.86964560 -0.03980597
## [3,] -0.1535415  0.86964560  1.00000000 -0.05833673
## [4,]  0.1416466 -0.03980597 -0.05833673  1.00000000
```

PercentMarried and PctMarriedHouseholds are clearly highly correlated with correlation value 0.87.

We can further check using Pearson Correlation Coefficient test. #Correlation Test

```
##
##  Pearson's product-moment correlation
##
## data:  cancer2$PercentMarried and cancer2$PctMarriedHouseholds
## t = 93.811, df = 2836, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8603819 0.8783348
## sample estimates:
##       cor
## 0.8696456
```

Multicollinearity for Employment There is strong correlation between PctUnemployed16_over and PctEmployed16_Over. We run a correlation test between these two variables to check to see if there is an yevidence of multi colinearity that should be investigated later.

```
##
##  Pearson's product-moment correlation
##
## data:  PctUnemployed16_Over and PctEmployed16_Over
## t = -45.251, df = 2836, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6683964 -0.6256390
## sample estimates:
##       cor
## -0.6475271
```

Multicollinearity for Poverty Poverty with PctUnemployed_Over16, PctEmployed_Over16, PctPrivateCoverage, PctEmpPrivCoverage, PctPublicCoverage

```
##       povertyPercent   PctEmployed16_Over PctUnemployed16_Over
##                 1.00                -0.74                 0.65
##   PctPrivateCoverage   PctEmpPrivCoverage    PctPublicCoverage
##                -0.82                -0.68                 0.65
```

Correlation is at least 0.65. In particular -0.74 with PctEmployed16_Over, -0.82 with PctPrivate-Coverage.

Proceed using Pearson Correlation Coefficent test. We get p-values 2.2e-16 for all five tests. We can conclude povertyPercent is significant correlated to these five variables.

```
##
##  Pearson's product-moment correlation
##
## data:  cancer2$povertyPercent and cancer2$PctEmployed16_Over
## t = -58.073, df = 2836, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7533906 -0.7197517
## sample estimates:
##        cor
## -0.7370272

##
##  Pearson's product-moment correlation
##
## data:  cancer2$povertyPercent and cancer2$PctUnemployed16_Over
## t = 46.131, df = 2836, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6332117 0.6752769
## sample estimates:
##      cor
## 0.654751

##
##  Pearson's product-moment correlation
##
## data:  cancer2$povertyPercent and cancer2$PctPrivateCoverage
## t = -77.209, df = 2836, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8346941 -0.8109497
## sample estimates:
##        cor
## -0.8231815
```

```
##
##  Pearson's product-moment correlation
##
## data:  cancer2$povertyPercent and cancer2$PctEmpPrivCoverage
## t = -49.589, df = 2836, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7006974 -0.6612594
## sample estimates:
##       cor
## -0.6814728
```

```
##
##  Pearson's product-moment correlation
##
## data:  cancer2$povertyPercent and cancer2$PctPublicCoverage
## t = 45.734, df = 2836, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6298182 0.6721944
## sample estimates:
##       cor
## 0.6515142
```

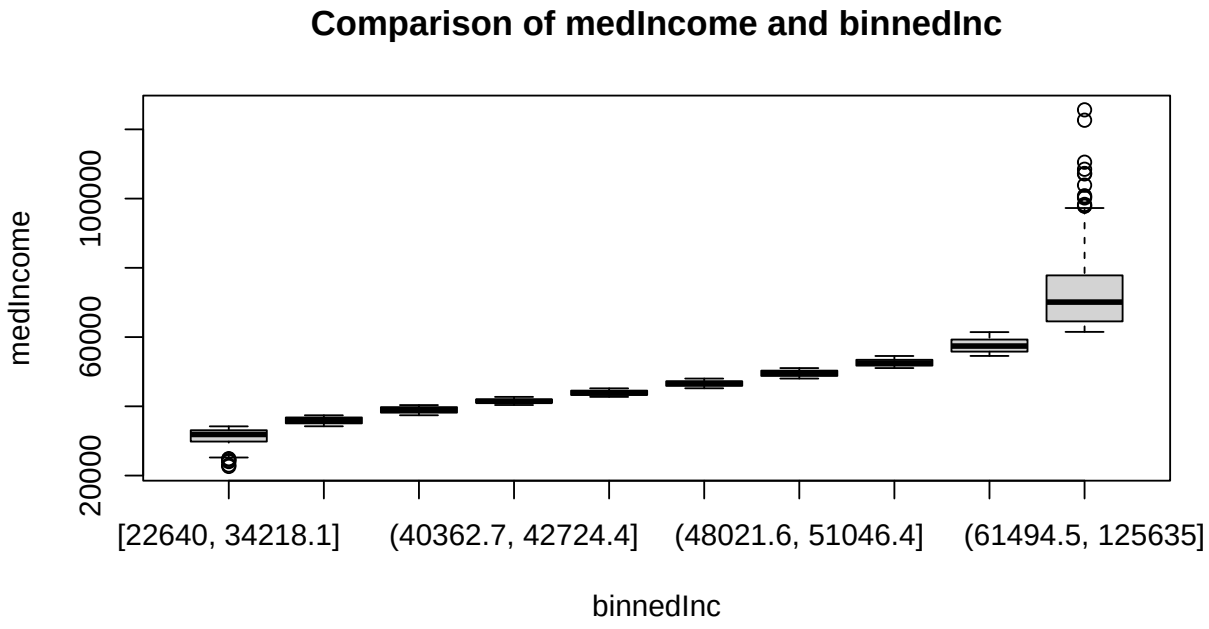


Figure 8: medIncome and binnedInc show very similar results. We will therefore consider only using medIncome in our model

Multicollinearity for Income We can observe that for the first 9 bins medIncome and binnedInc show very similar results. We will therefore consider only using medIncome in our model.

```
##
##  Pearson's product-moment correlation
##
## data:  cancer2$MedianAgeMale and cancer2$MedianAgeFemale
## t = 137.96, df = 2836, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9279690 0.9375231
## sample estimates:
##     cor
## 0.93291
```

As the p value is below 0.05 we can assume the correlation is significant and multicollinearity exists between the Median age Male and Median age Female.