Checking the summary and initial EDA

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.

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.

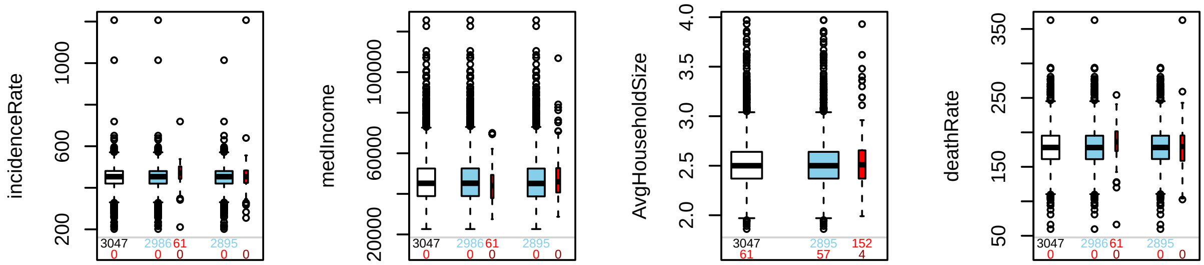


Figure 1: Pbox plots showing difference between missing and non-missing data

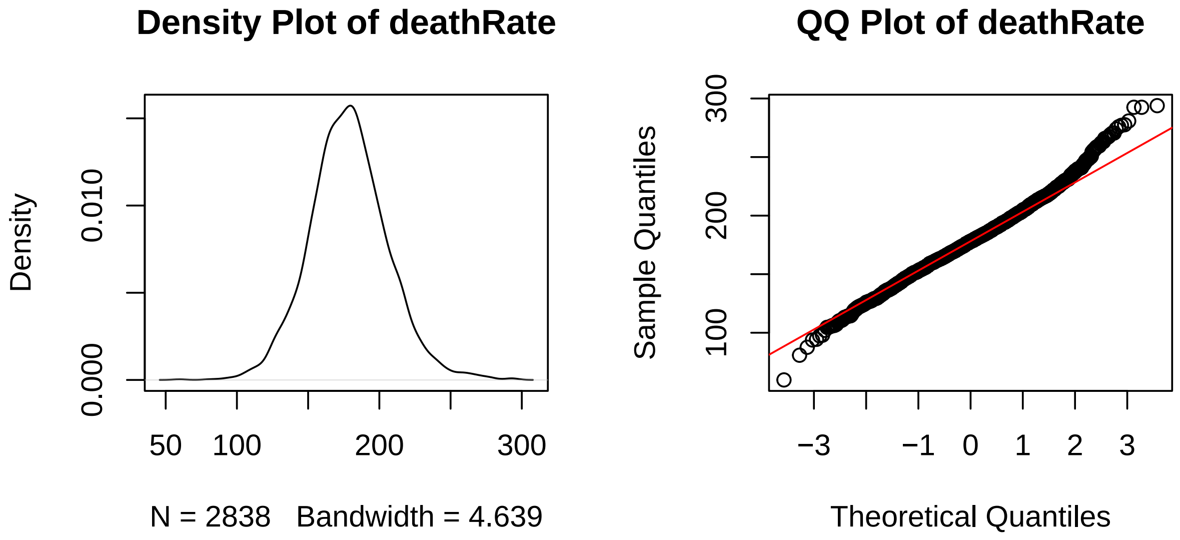
We use the pbox() function from the VIM package to check what these missing values represent.

From the above plots we note that the box plots with the missing data do not look significantly different from those without. The pbox plots for the other variables look similar to this so we conclude that the data that is missing is MCAR.

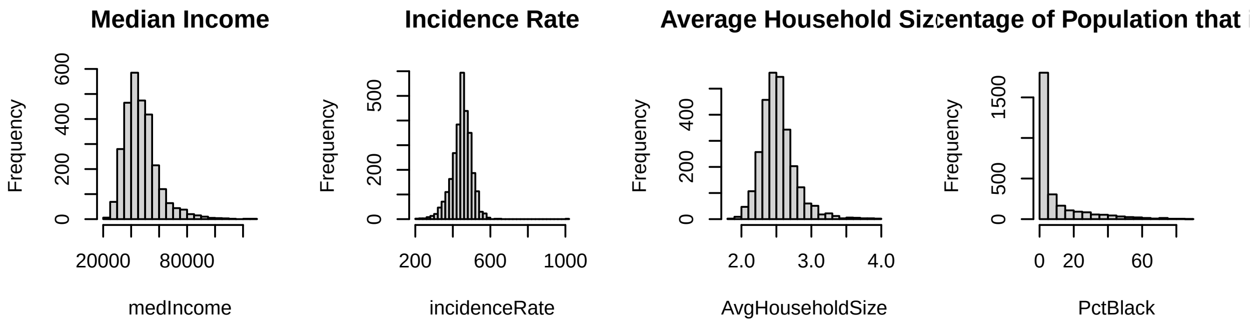
With our data we could replace all the data with an alternate source but as the proportion of missing data points is so small and it is MCAR it is safe to just remove the rows with missing data from our data set.

Analysis of deathRate

Our assignment brief tells us we should investigate deathRate as a repsonse variable when it comes to our investigation. So we first make sure that a normal linear model is appropriate by making sure that deathRate is normally distributed.

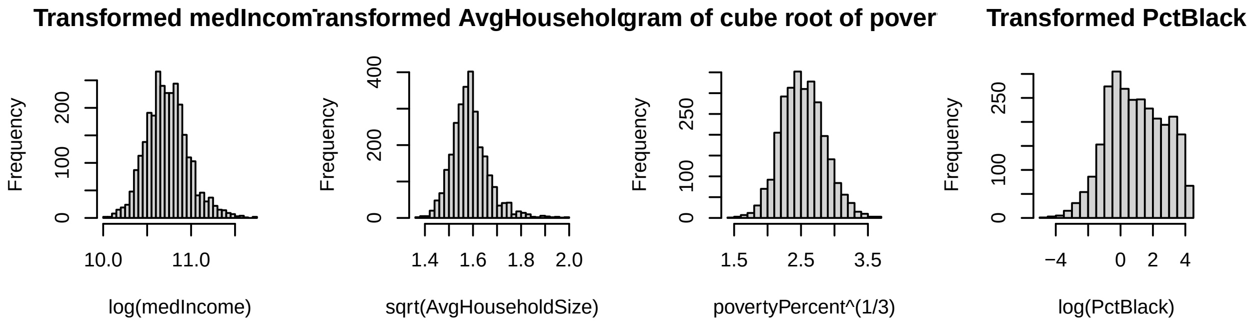


Histograms



Analysis on PctBlack, MedIncome, AverageHousehold, Incidence Rate 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. 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.



Analysis of Transformed Histograms 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

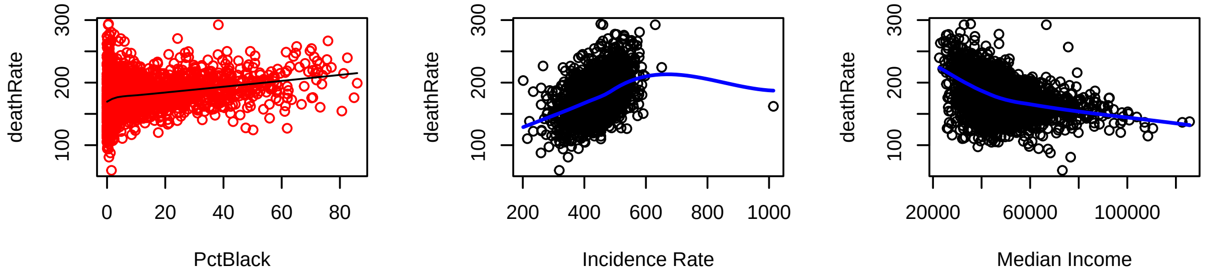


Figure 2: Plots showing deathRate against other variables

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.

There is definite heteroscedasticity in medIncome and for both MedianAgeFemale and MedianAgeMale 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 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

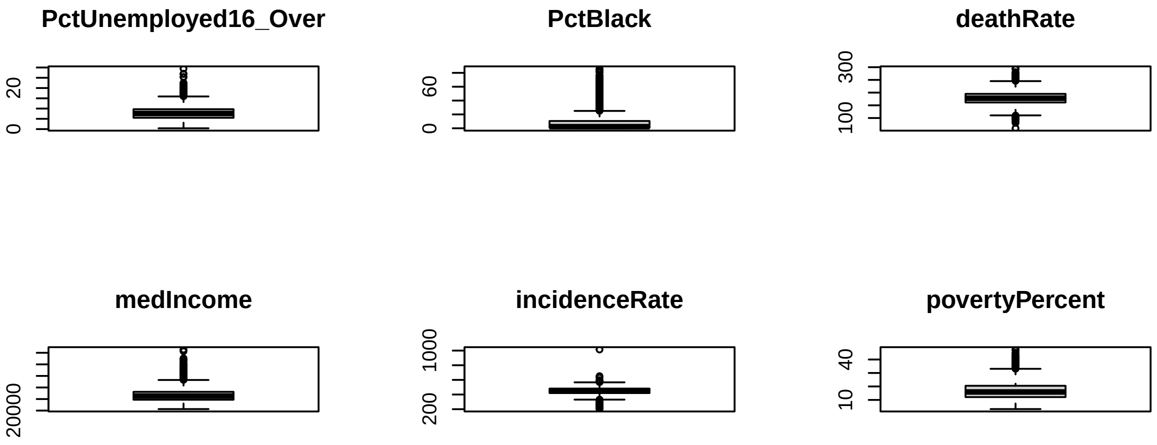


Figure 3: BoxPlots of our variables

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.

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 From the plot we can see that there is potential multicollinearity between: PercentMarried and PctMarried Households (correlation 0.87), PctUnemployed16_over and PctEmployed16_Over (correlation …), Poverty with PctEmployed_Over16 (-0.74) and PctPrivateCoverage (-0.82),

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

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.

Correlation Tests on the above Variables