Exploratory Data Analysis on the US Cancer Dataset

Written by

Alex Walters, 1921659 Jena Moteea, 1939940 Remus Gong, 1918934
James Keith, 1827052

Contents

1 Executive summary

- **Missing values have been identified to likely be missing completely at random (MCAR) and therefore we suggest that it can be removed. Some data for the AvgHouseholdSize data is likely to be incorrect and also MCAR so should be removed. Overall, the proportion of missing values in all variables is small and so it is not a serious problem.**
- **Some variables, namely PctBlack, incidenceRate and medIncome, should be further investigated and altered to further increase the strength of our analysis before building a statistical model.**
- **Some variables show very similar data and therefore some variables will not be significant in building a statistical model. We suggest omitting binnedInc due to high correlation with medIncome and either PercentMarried or PctMarried-Households due to high correlation between variables. Further analysis should be provided to check any outliers' nature and more rigorously check similarities between variables before building a statistical model.**

2 Findings

This report will be an investigation into the distributions of the variables and the relationships between them in the cancer dataset in US counties, where the data is adapted from Noah Rippner (Rippner (2016)). We will conduct exploratory data analysis which will leave us with the opportunity to create a statistical model after our research.

2.1 Initial EDA and missing values

The character variable Geography is just an identifier of the observation and can be ignored for statistical analysis and should not be used in a linear model. However, it can be utilised for data visualisation and analysis of geographic trends in the United States. We identified missing values in PctEmployed16_Over and abnormal values in AvgHouseholdSize. The best course of action is to remove the observations associated with them before further analysis of the data as they are likely missing completely at random.

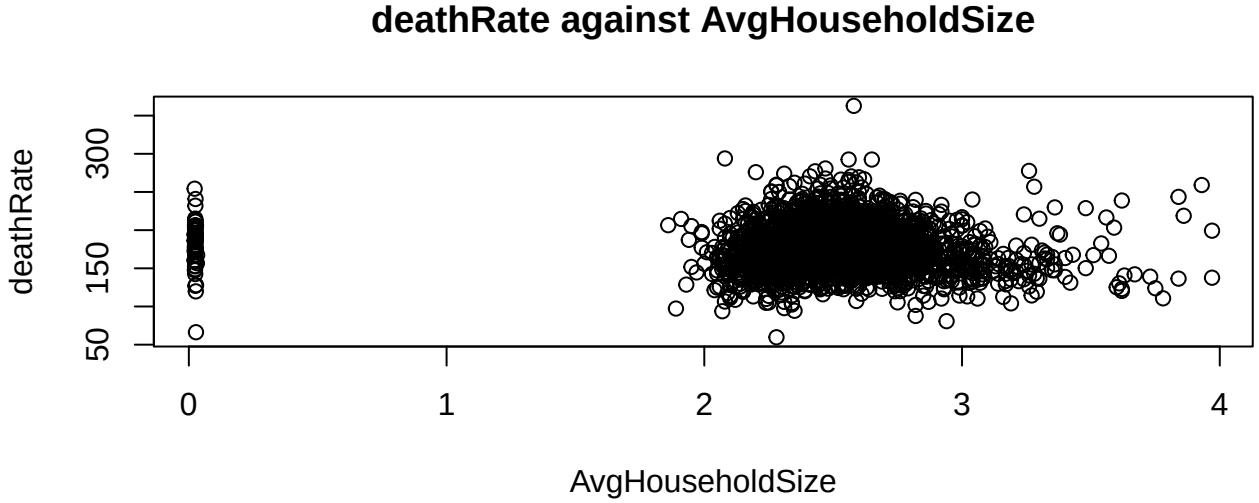


Figure 1: The abnormal values near zero were investigated and removed before further analysis

2.2 Skewness, Heteroscedasticity and Linearity

It appears most predictor variables are linear and there are no signs of heteroscedasticity. However, we identified heteroscedasticity in three variables: PctBlack, incidenceRate, medIncome. Potential outliers may influence the heteroscedascity. We recommend using power transformations to fix heteroscedasticity. There exists non-linearity in AvgHouseholdsize, MedianAgeFemale and MedianAgeMale. We suggest using a more complex model to improve linearity. Some variables, most notably PercentMarried and medIncome, are skewed and we tested simple power transformations to fix the skewness, which proved effective. However, due to a massive right skew in PctBlack and the data for this variable being non-normal, we could not completely remove the skewness and so this needs to be further explored when fitting a model. We suggest using a spread level plot and power transformations to fix heteroscedasticity. We noticed that the power transformations used in fixing heteroscedasticity and skewness are different and when building a model heteroscedasticity should be prioritised to issues in normality.

2.3 Outliers

We used box plots to discover outliers. All variables have outliers in our data, but we identified a considerable number of outliers in medIncome, PctBlack and incidenceRate. The outliers in medIncome and PctBlack are reasonable after some research of the US demographics (review (2022)). However, there are some severe outliers in incidenceRate. We could further investigate into this by using Cook's distance before we can make decisions on how we can treat the outliers.

2.4 Multicollinearity

There are many variables that suffer from multicollinearity due to representing very similar things. For example, PercentMarried and PctMarriedHouseholds highly correlated and could be considered to represent the same force. Most notably, binnedInc and medIncome represent very similar data, we suggest removing binnedInc from the model or including an interaction term between the two variables. Multicollinearity is also true for many other variables in the dataset, and we should investigate variance inflation factors (VIF) and other multicollinearity diagnosis methods once a model has been fitted. If our initial suspicions about multicollinearity are proven true by further analysis, then we can consider removing one of the co-linear variables to reduce model complexity.

2.5 Correlation with deathRate

We found that incidenceRate, medIncome, povertyPercent, PctEmployed16_Over, PctUnemployed16_Over, PctPrivateCoverage and PctPublicCoverage are highly correlated with deathRate and should be taken into consideration as important predictors for when it comes to constructing a linear model. There are some predictors that have almost no correlation at all such as AvgHouseholdSize and MedianAgeFemale. Initially, it may be expected that counties with older populations will incur more deaths, this is not represented by our data and so we would need to investigate this further once we've made a linear model.

2.6 Recommendations

We suggest the following recommendations prior to fitting a linear model:

- Remove observations with likely MCAR values and incorrect values.
- Apply appropriate transformations, using a spread level plot and power transformations, to fix cases of non-linearity, heteroscedasticity, skewness, and non-normality to satisfy the model assumptions.
- Further investigate outliers, using Cook's distance, to ensure they do not have undue influence when we build a model.
- Perform further analysis, using Pearson's correlation coefficient and VIF, on highly correlated predictor variables to check for multicollinearity.
- Prioritise the variables with significant correlation with deathRate in the model.

3 Statistical Methodology

3.1 Checking the summary and initial EDA

Looking at the dataset we note that there is one character variable, one factor variable and sixteen continuous variables (See Appendix (B) 1.1).

There are 3047 observations of data in our dataset. That is a large amount of data, but it doesn't actually equal the total amount of US counties which is 3142 in total (Sawe (2018)). The proportion of counties in our data is large enough so that the data can still serve as a good indicator.

There are 152 missing values in PctEmployed16_Over which need to be checked. There are 61 values in AvgHouseholdSize below 0.1 which should be investigated before further analysis due to being severely small (See Appendix (B) 1.2).

We identify one of these points and investigate it:

Table 1: Berkeley County's AvgHouseholdSize

Geography	AvgHouseholdSize
Berkeley County, West Virginia	0.0263

To check the validity of this data point we find an alternate source of the data (Bureau (2013)).

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 missing completely at random (MCAR).

3.2 Missing Value Exploration

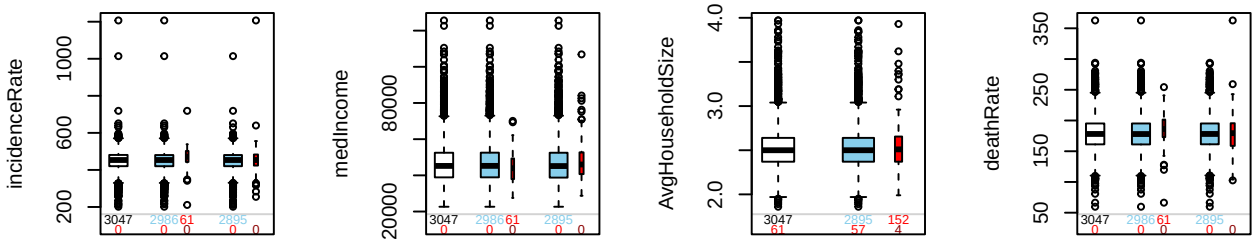


Figure 2: Box Plots showing the difference between missing and non-missing data in four variables

We use the pbox() function from the VIM package to check what these missing values represent (Kowarik and Templ (2016)). From the plots (Fig. 2 and Appendix (B) 1.3) we note that the box plots with the missing data do not look significantly different from those without missing data. The box plots for the other variables look similar to this which suggests that the data that is missing is likely MCAR for both AvgHouseholdSize and PctEmployed16_Over.

As the proportion of missing data points is relatively small and the values are likely to be MCAR, the problem of missing values is not that serious. It should be safe to remove the rows with missing

data from our data set. This will not make the data any less representative and shouldn't affect our statistical analysis when we come to build a linear model, other than slightly increasing the standard error. From now on we consider the dataset without observations which contain missing and incorrectly inputted values.

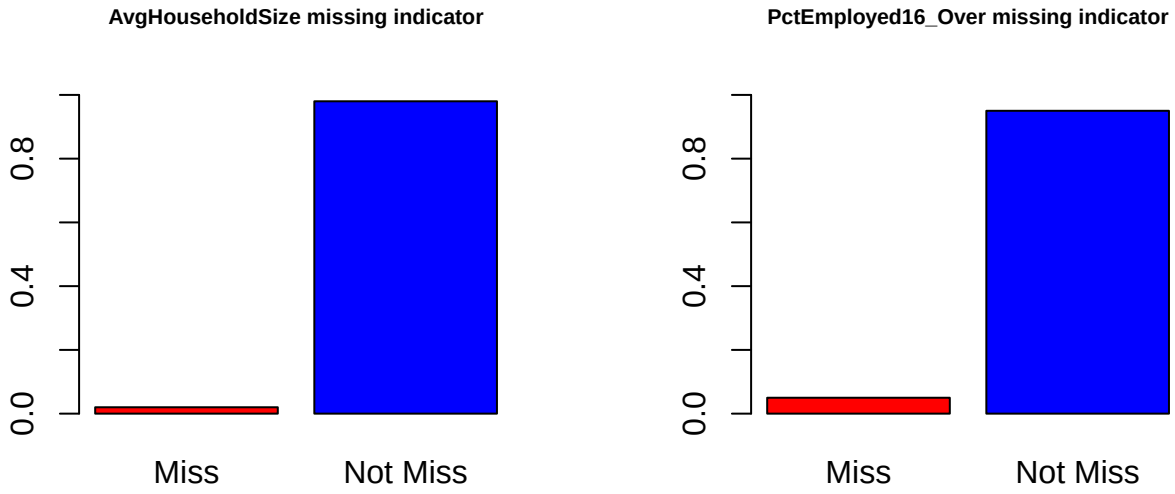


Figure 3: The proportion of missing values are low

3.3 Univariate Plots

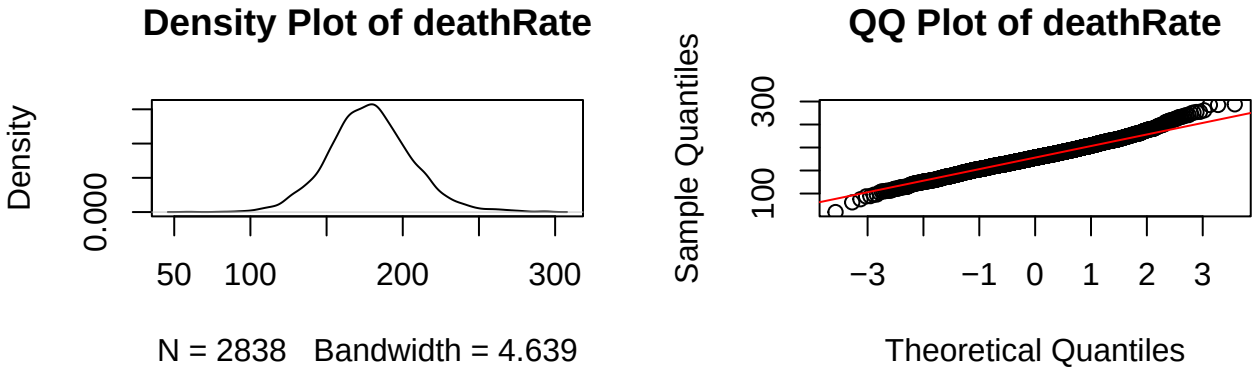


Figure 4: Density and QQ-Plots of deathRate

We should investigate deathRate as a response variable. We first test if a normal linear model is appropriate by making sure that deathRate is normally distributed. From deathRate's density and QQ Plots (Fig. 4) we can see that the variable deathRate is normally distributed and so a normal linear model is appropriate to use.

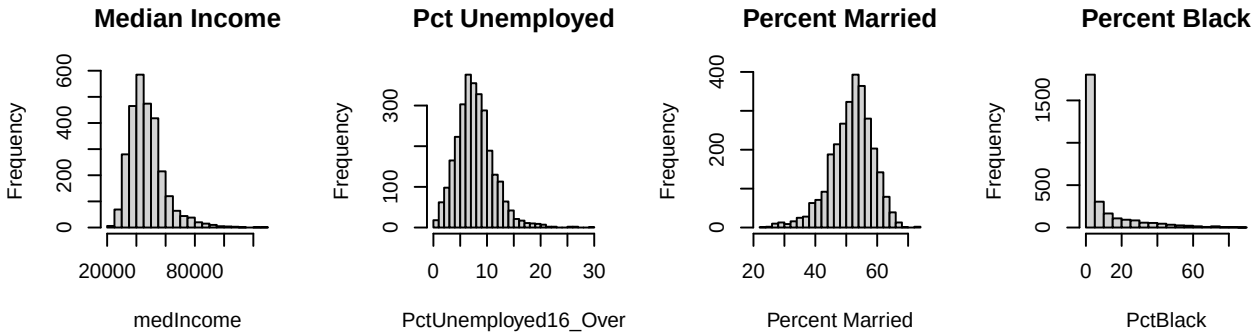


Figure 5: Histograms of four skewed predictor variables

The 4 plots above (Fig. 5) represent the most skewed variables and hence indicate non-normality. Most of our predictor variables look normally distributed from their histograms (see Appendix (B) 2.1) but medIncome, PctUnemployed16_Over, PercentMarried, PctMarriedHouseholds, povertyPercent and PctBlack all have significant skew. Transformations, such as log or square root, of skewed variables should be considered to fix skewness as seen in Fig. 4.

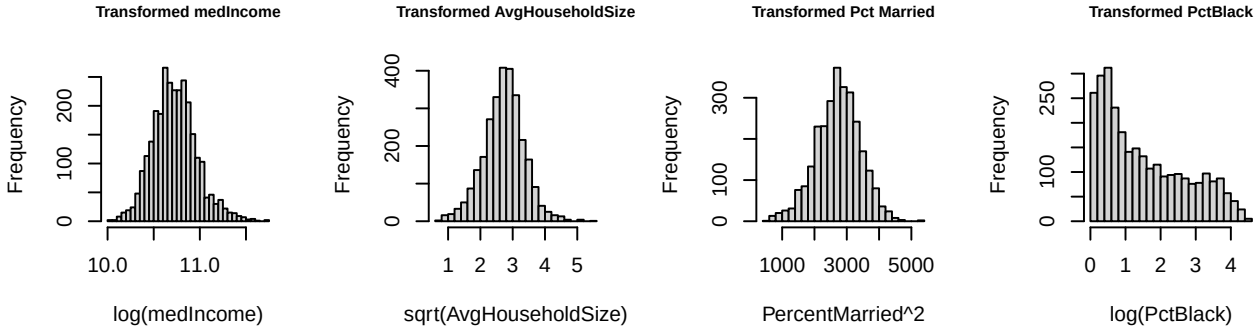


Figure 6: Histograms of four transformed variables previously skewed

From the above (Fig. 6) we note that some simple transformations can be applied to fix most of these variables' skew (log transform for median income for large right skew, square root for AvgHouseholdSize for slight right skew, square transform for PercentMarried for slight left skew). For those three variables the skew and normality are mostly fixed. However, for PctBlack a log transform is not sufficient to fix the large right skew. A shift in pctBlack data before log transforming it is necessary as there exists some zero-value data in pctBlack, which cannot be log transformed. This indicates that the data may not be normally distributed and would need to be handled differently when it comes to our statistical model.

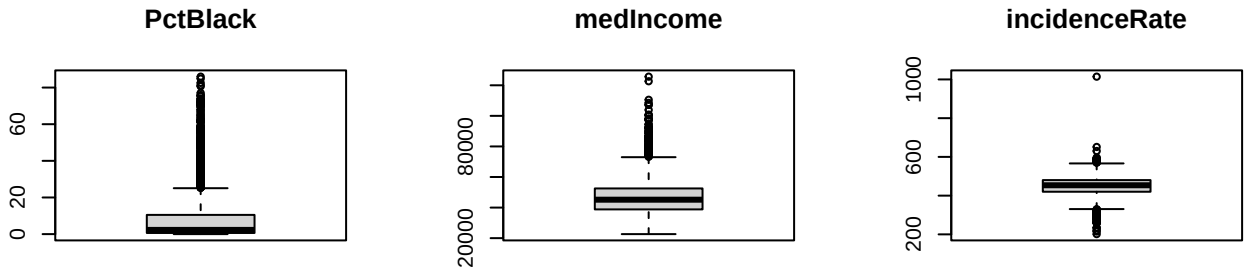


Figure 7: BoxPlots of PctBlack, medIncome, and incidenceRate

Our box plots (Fig. 7 and see Appendix (B) 2.3) show that we have a large number of outliers across all our variables apart from binnedInc, which would be impossible due to the bins intervals. We have a severe number of outliers in medIncome. This is most likely due to natural circumstances such as a CEO of a large company or a doctor earning a higher salary (Team (2021)). We also observe significant outliers in PctBlack and this is illustrated by the very long tail as shown in the histogram above (See Fig. 4). This might be due to PctBlack being an unstable predictor variable. We observe significantly high percentages of over 50% in the south and southeast region of the US, in particular, in Mississippi, Georgia, Alabama, and North and South Carolina. They form part of the top 10 US states with the highest percentage of Black residents. (review (2022))

The boxplot for incidence rate shows the existence of extreme high values which is also illustrated in our bivariate plots (Fig. 6). There are potential outliers in PctPrivateCoverage and povertyPercent. We might want to further investigate into these, using Cook's distance, and decide how we might want to treat them before fitting the model. Possible options might include removing the outliers or imputing them.

We suggest that the incidence rate in Williamsburg city, Virginia can be considered as unusual and should be further analysed once a model is fitted (ghrconnects.org (2022)).

3.4 Bivariate Plots

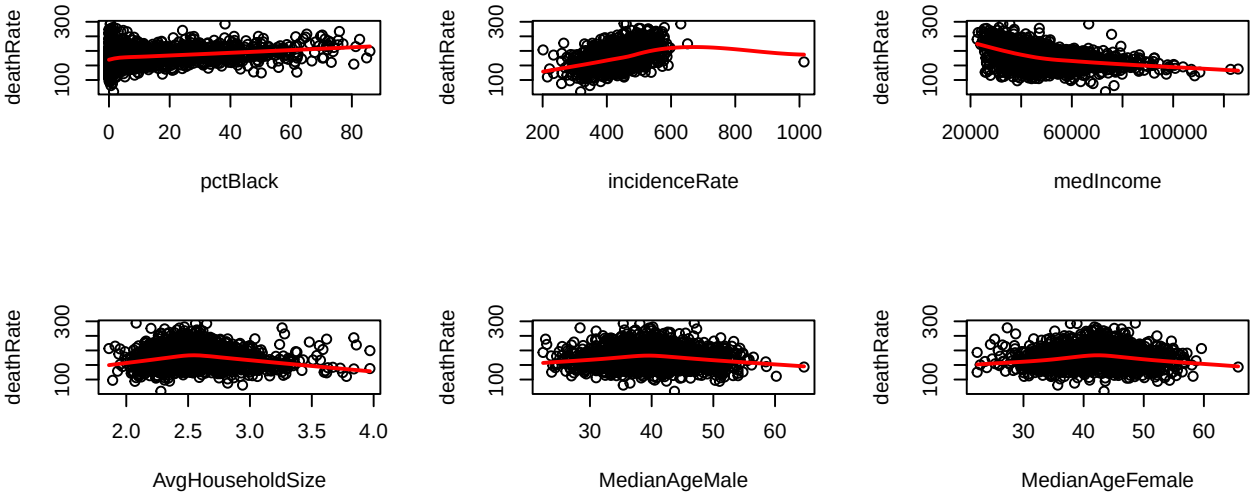


Figure 8: Plots showing deathRate against predictor variables

Most predictor variables in our dataset show signs of linearity and no heteroscedasticity (See Appendix (B) 3.1). However, from the bivariate plots above, we can observe heteroscedasticity in PctBlack, incidenceRate and medIncome. We might need to perform further investigation after fitting a model and we can use spreadLevelPlot() to find an appropriate power transformation to fix heteroscedasticity. We were also able to observe unusual outliers of incidence rate that might have a high influence explaining its heteroscedasticity and non-linearity. The transformations that we have tested to fix the skew of some variables may not be the same as those that would fix heteroscedasticity or non-linearity. In this case we should prioritise the homoscedasticity and linearity model assumptions as those are more important to fitting a good model than normality.

Moreover, we can see non-linearity in AvgHouseholdsize, MedianAgeFemale and MedianAgeMale (Fig. 6). We notice a concave shape for incidence rate and non-monotonic data, as for AvgHouseholdsize, so we advise potentially testing a quadratic term to improve linearity following further analysis such as residual plots.

3.5 Multicollinearity

We can see that there is potential multicollinearity between: PercentMarried and PctMarried Households (correlation 0.87), PctUnemployed16_over and PctEmployed16_Over (correlation -0.65), and MedianAgeFemale and MedianAgeMale (correlation 0.94) (See Appendix (B) 4.1). We further used Pearson correlation test (See Appendix (B) 3.2) to check for multicollinearity between percentPoverty and PctEmployed_Over16 (correlation -0.74), PctPrivateCoverage (correlation -0.82), PctEmpPrivCoverage (correlation -0.68), and PctPublicCoverage (correlation 0.65). Therefore, we might consider removing some of the predictor variables due to co-llinearity, as one of the two co-linear variables will have little significance in the model. Moreover, for the first 9 bins medIncome and binnedInc show very similar results (See Appendix (B) 3.2) we suggest removing binnedInc from the model, or we could also consider using an interaction term between the two variables.

3.6 Correlation with deathRate

We note that none of the variables highly correlate with deathRate, but there are several variables that weakly correlate with deathRate that should be noted when it comes to building a model. These are incidenceRate, medIncome, povertyPercent, PctEmployed16_Over, PctUnemployed16_Over, PctPrivateCoverage and PctPublicCoverage. However, a lot of these variables also highly correlate with each other so the correlation with deathRate will result from these variables measuring very similar forces. It should also be noted that MedianAgeMale, MedianAgeFemale and AvgHouseholdSize have extremely weak correlation with deathRate, so they should have lower priority when constructing a linear model.

Table 2: Correlations of variables with deathRate

incidenceRate	medIncome	povertyPercent	PctEmployed16_Over	PctPrivateCoverage	PctPublicCoverage	MedianAgeFemale	AvgHouseholdSize
0.42	-0.43	0.43	-0.42	-0.39	0.4	0.01	-0.04

4 Author's Contributions

Throughout this project Alex Walters contributed towards formatting the report, analysis on missing values, the statistical methodology, key findings, appendix, presentation slides and references. Jena Moteea contributed towards the analysis on multicollinearity, statistical methodology, key findings, references, appendix, and the presentation slides. James Keith contributed towards structuring the report with the initial analysis, statistical methodology, key findings, appendix, the presentation slides, and references. Remus Gong contributed towards the analysis on skewness, statistical methodology, key findings, executive summary, appendix, and the presentation slides.

5 References

Bureau, United States Census. 2013. "American Community Survey, S1101 HOUSEHOLDS AND FAMILIES." United States Census Bureau. https://data.census.gov/cedsci/table?q=average%20household%20size&g=0500000US54003&y=2013&tid=ACSST1Y2013.S1101.

ghrconnects.org. 2022. "Community Indicators Dashboard, Cancer Incidence Rate." Greater Hampton Roads. https://www.ghrconnects.org/indicators/index/view?indicatorId=162&localeId=3004&localeFilterId=49&localeChartIdxs=1%7C2%7C3.

Kowarik, Alexander, and Matthias Templ. 2016. "Imputation with the R Package VIM." *Journal of Statistical Software*. https://doi.org/10.18637/jss.v074.i07.

review, World population. 2022. "Black Population by State 2022." U.S. Census Bureau. https://worldpopulationreview.com/state-rankings/black-population-by-state.

Rippner, Noah. 2016. "OLS Regression Challenge." Noah Rippner. https://data.world/nrippner/ols-regression-challenge.

Sawe, Benjamin Elisha. 2018. "How Many Counties Are in the United States?" World Atlas. https://www.worldatlas.com/articles/how-many-counties-are-in-the-united-states.html.

Team, Indeed Editorial. 2021. "Top 100 Highest Paying Jobs." Indeed. https://www.indeed.com/career-advice/finding-a-job/top-100-highest-paying-jobs.

6 Appendix

6.1 Appendix A: Code used in report

2.1: Initial EDA and missing values

```
plot(deathRate ~ AvgHouseholdSize, data = cancer,
     main = "deathRate against AvgHouseholdSize")
```

3.1: Checking the summary and initial EDA

```
kable(subset(cancer, AvgHouseholdSize < 0.1)[1,c("Geography", "AvgHouseholdSize")],
      align = "l", caption = "Berkeley County's AvgHouseholdSize")
```

```
#Makes a dataset where the abnormally low values in AvgHouseholdSize are NA
cancer1 <- cancer
cancer1$AvgHouseholdSize[which(cancer1$AvgHouseholdSize < 0.1)] <- NA
```

3.2: Missing value exploration

```
par(mfrow = c(1,4))
for (i in c(2,3,8,18)){
  pbox(cancer1,pos=i)
}
```

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

3.3: Univariate Plots

```
par(mfrow = c(1,2))
plot(density(cancer2$deathRate), main = "Density Plot of deathRate")
qqnorm(cancer2$deathRate, main = "QQ Plot of deathRate")
qqline(cancer2$deathRate, col = "red")
```

```
par(mfrow = c(1,4))
hist(cancer2$medIncome, breaks = 30, main = "Median Income",
     ylab = "Frequency", xlab = "medIncome")
hist(cancer2$PctUnemployed16_Over, breaks = 30, main = "Pct Unemployed",
     ylab = "Frequency", xlab = "PctUnemployed16_Over")
hist(cancer2$PercentMarried, breaks = 30, main = "Percent Married",
     ylab = "Frequency", xlab = "Percent Married")
hist(cancer2$PctBlack, breaks = 30, main = "Percent Black",
     ylab = "Frequency", xlab = "PctBlack")
```

```
par(mfrow = c(1,4))
hist(log(cancer2$medIncome), breaks = 30, main = "Transformed medIncome",
     ylab = "Frequency", xlab = "log(medIncome)", cex.main = 0.8)
hist(sqrt(cancer2$PctUnemployed16_Over), breaks = 30,
main = "Transformed AvgHouseholdSize",
     ylab = "Frequency",
     xlab = "sqrt(AvgHouseholdSize)", cex.main = 0.8)
hist(cancer2$PercentMarried^2, breaks = 30, main = "Transformed Pct Married",
```

```
     ylab = "Frequency", xlab = "PercentMarried^2", cex.main = 0.8)
hist(log(cancer2$PctBlack + 1), breaks = 30, main = "Transformed PctBlack",
     ylab = "Frequency", xlab = "log(PctBlack)", cex.main = 0.8)
```

3.4: Bivariate Plots

```
par(mfrow = c(2,3))
with(cancer2, scatter.smooth(deathRate~PctBlack, ylab = "deathRate",
                     xlab = "pctBlack", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~incidenceRate, ylab = "deathRate",
                     xlab = "incidenceRate", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~medIncome, ylab = "deathRate",
                     xlab = "medIncome", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~AvgHouseholdSize, ylab = "deathRate",
                     xlab = "AvgHouseholdSize", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~MedianAgeMale, ylab = "deathRate",
                     xlab = "MedianAgeMale", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~MedianAgeFemale, ylab = "deathRate",
                     xlab = "MedianAgeFemale", lpars = list(col = "red", lwd = 2)))
```

```
par(mfrow = c(1,3))
with(cancer2, boxplot(PctBlack, main = "PctBlack"))
with(cancer2, boxplot(medIncome, main = "medIncome"))
with(cancer2, boxplot(incidenceRate, main = "incidenceRate"))
```

3.5: Correlation with deathRate

```
kable(round(cor(cancer2$deathRate, cancer2[,c(2,3,5,10,12,14,7,8)]), digits = 2),
      align = "l", caption = "Correlations of variables with deathRate")
```

6.2 Appendix B: Code not included in the report

1: Summary of the Data and initial EDA

1.1: Summary

```
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.: 38882   (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
##
```

```
# Checking the summary output, can spot missing values in PctEmployed16_Over
# The minimum value in AvgHouseholdSize is very small and should be investigated.
```

1.2: Dataset Length

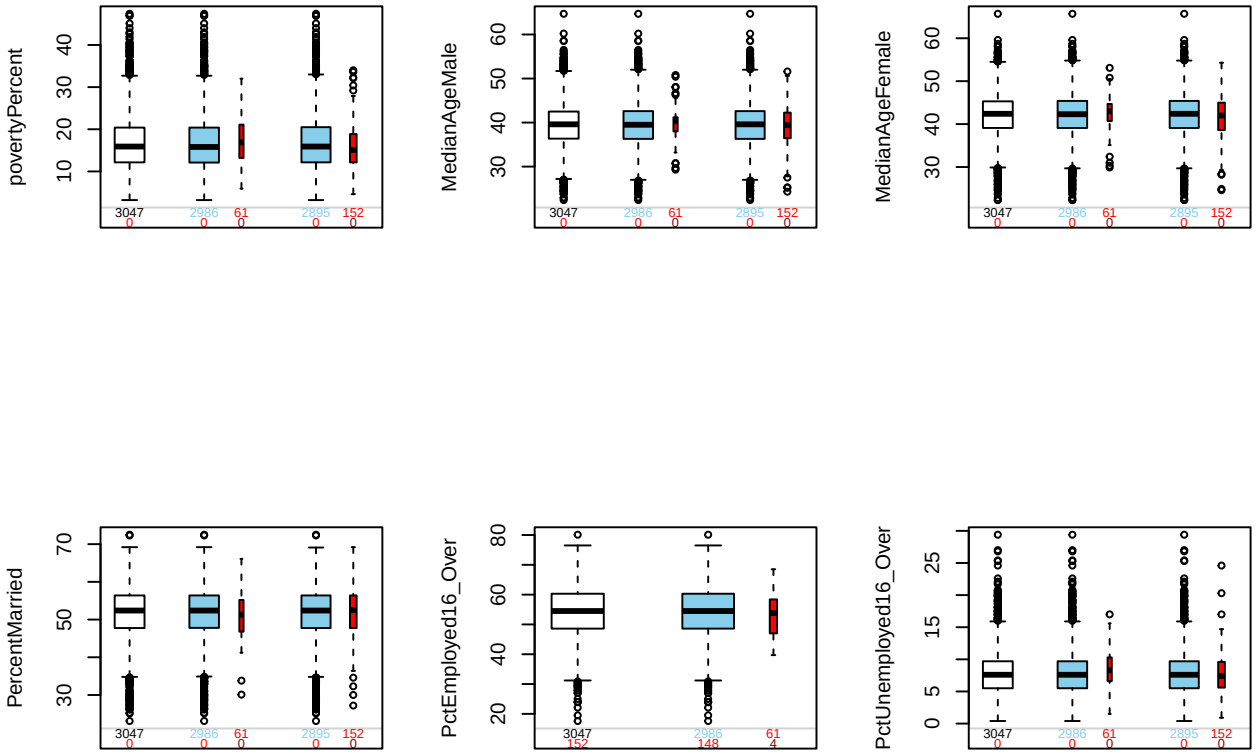
```
#Checking number of observations and suspiciously low values
length(cancer$Geography)
length(cancer$AvgHouseholdSize[cancer$AvgHouseholdSize<0.1])
```

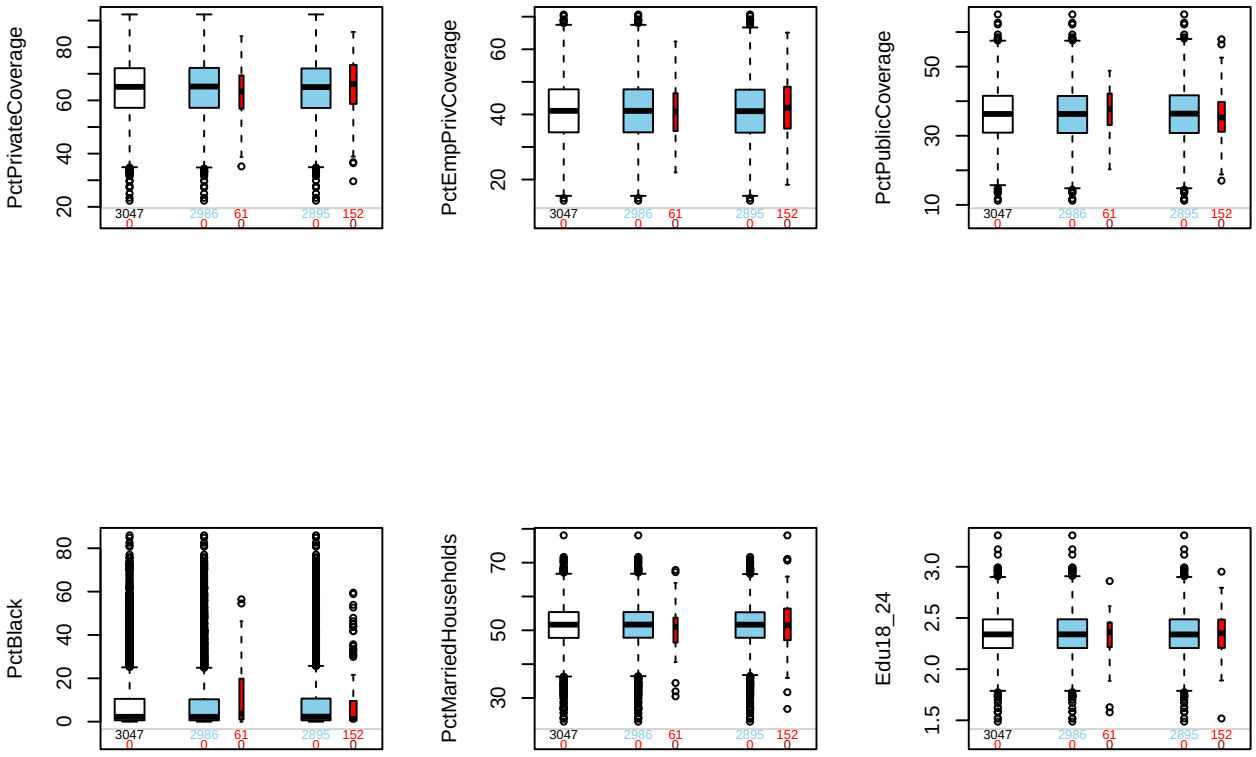
```
## [1] 3047
## [1] 61
```

1.3: Missing value parallel boxplots

```
# Check if data is skewed without missing values
par(mfrow = c(2,3))
```

```
for (i in c(5:7, 9:11)){
  pbox(cancer1,pos=i)
}
for (i in 12:17){
  pbox(cancer1,pos=i)
}
```



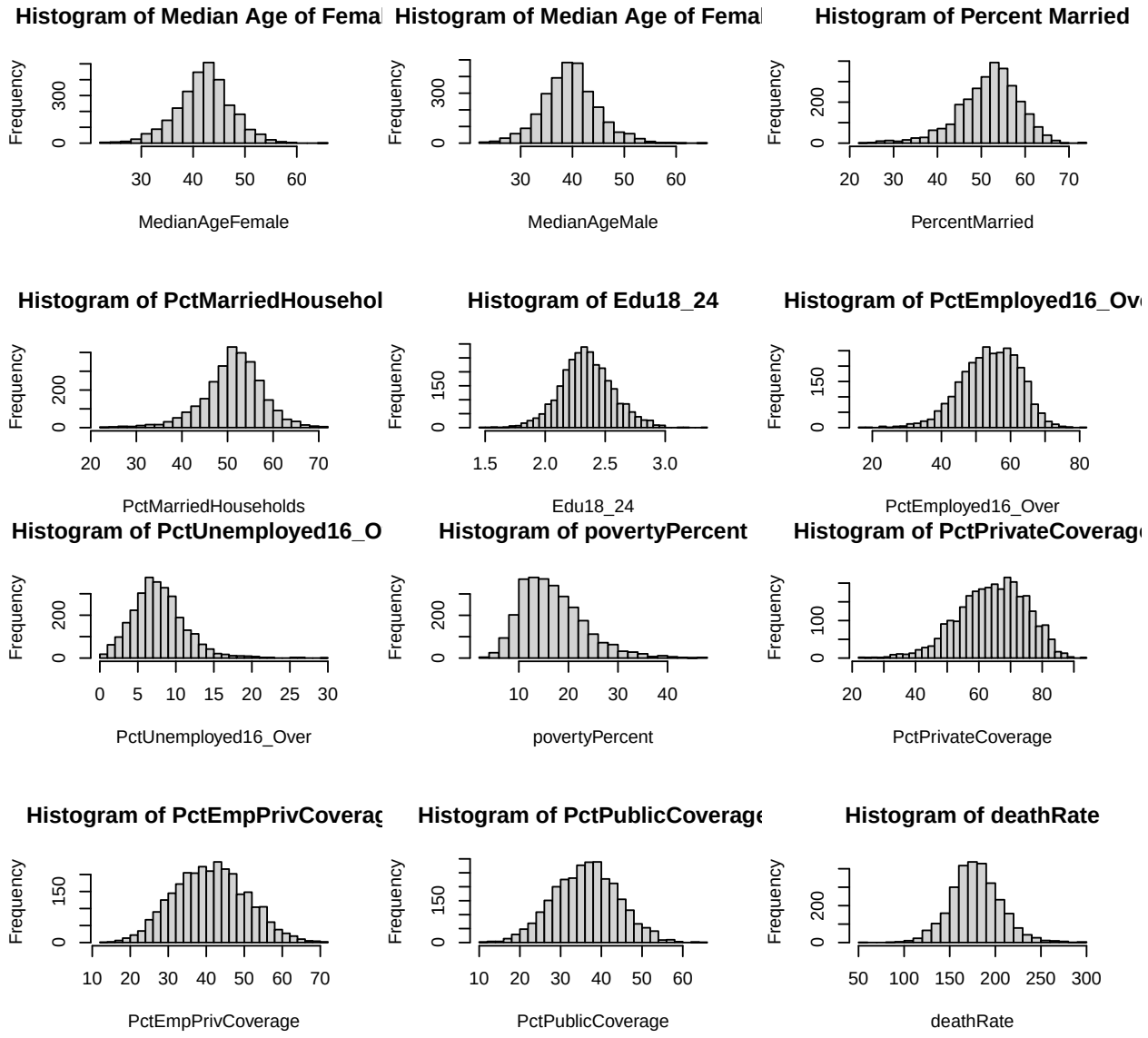


2: Univariate Plots

2.1: Histograms

```
# Checking skewness using histograms of numerical variables
par(mfrow = c(2,3))
with(cancer2, hist(MedianAgeFemale, breaks = 30,
main = "Histogram of Median Age of Females"))
with(cancer2, hist(MedianAgeMale, breaks = 30,
main = "Histogram of Median Age of Females"))
with(cancer2, hist(PercentMarried, breaks = 30,
main = "Histogram of Percent Married"))
with(cancer2, hist(PctMarriedHouseholds, breaks = 30,
main = "Histogram of PctMarriedHouseholds"))
with(cancer2, hist(Edu18_24, breaks = 30, main = "Histogram of Edu18_24"))
with(cancer2, hist(PctEmployed16_Over, breaks = 30,
main = "Histogram of PctEmployed16_Over"))
with(cancer2, hist(PctUnemployed16_Over, breaks = 30,
main = "Histogram of PctUnemployed16_Over"))
with(cancer2, hist(povertyPercent, breaks = 30,
main = "Histogram of povertyPercent"))
with(cancer2, hist(PctPrivateCoverage, breaks = 30,
main = "Histogram of PctPrivateCoverage"))
with(cancer2, hist(PctEmpPrivCoverage, breaks = 30,
main = "Histogram of PctEmpPrivCoverage"))
with(cancer2, hist(PctPublicCoverage, breaks = 30,
main = "Histogram of PctPublicCoverage"))
```

```
with(cancer2, hist(deathRate, breaks = 30, main = "Histogram of deathRate"))
```



2.2: Transformed Histograms

```
# Histograms of transformed predictor variables
par(mfrow = c(1,2))
hist(cancer2$PctPrivateCoverage^(2), breaks = 30,
main = "Histogram of square of PctPrivateCoverage")
with(cancer2, hist(sqrt(PctUnemployed16_Over),
main = "Transformed PctUnemployed16_Over"))
```

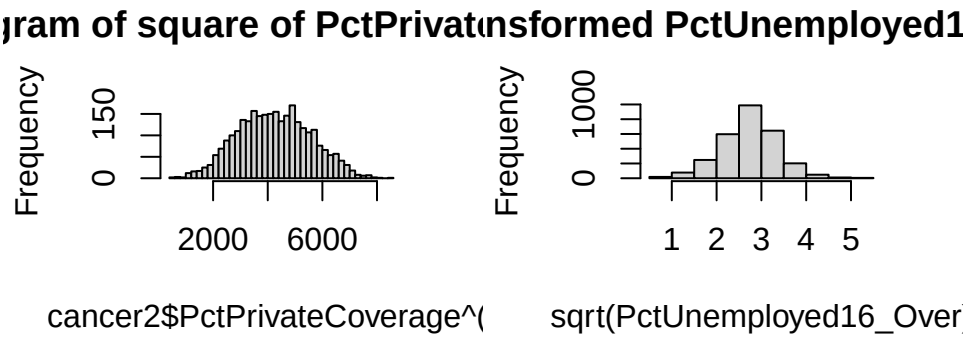


Figure 9: Our transformed histograms

2.3: Boxplots

```
# Checking outliers of numerical variables
par(mfrow = c(2,4))
with(cancer2, boxplot(PctUnemployed16_Over, main = "PctUnemployed16_Over"))
with(cancer2, boxplot(deathRate, main = "deathRate"))
with(cancer2, boxplot(povertyPercent, main = "povertyPercent"))
with(cancer2, boxplot(AvgHouseholdSize, main = "AvgHouseholdSize"))
with(cancer2, boxplot(PctEmployed16_Over, main = "PctEmployed16_Over"))
with(cancer2, boxplot(MedianAgeFemale, main = "MedianAgeFemale"))
with(cancer2, boxplot(MedianAgeMale, main = "MedianAgeMale"))
with(cancer2, boxplot(binnedInc, main = "binnedInc"))
with(cancer2, boxplot(PercentMarried, main = "PercentMarried"))
with(cancer2, boxplot(PctMarriedHouseholds, main = "PctMarriedHouseholds"))
with(cancer2, boxplot(Edu18_24, main = "Edu18_24"))
with(cancer2, boxplot(PctPrivateCoverage, main = "PctPrivateCoverage"))
with(cancer2, boxplot(PctEmpPrivCoverage, main = "PctEmpPrivCoverage"))
with(cancer2, boxplot(PctPublicCoverage, main = "PctPublicCoverage"))
```

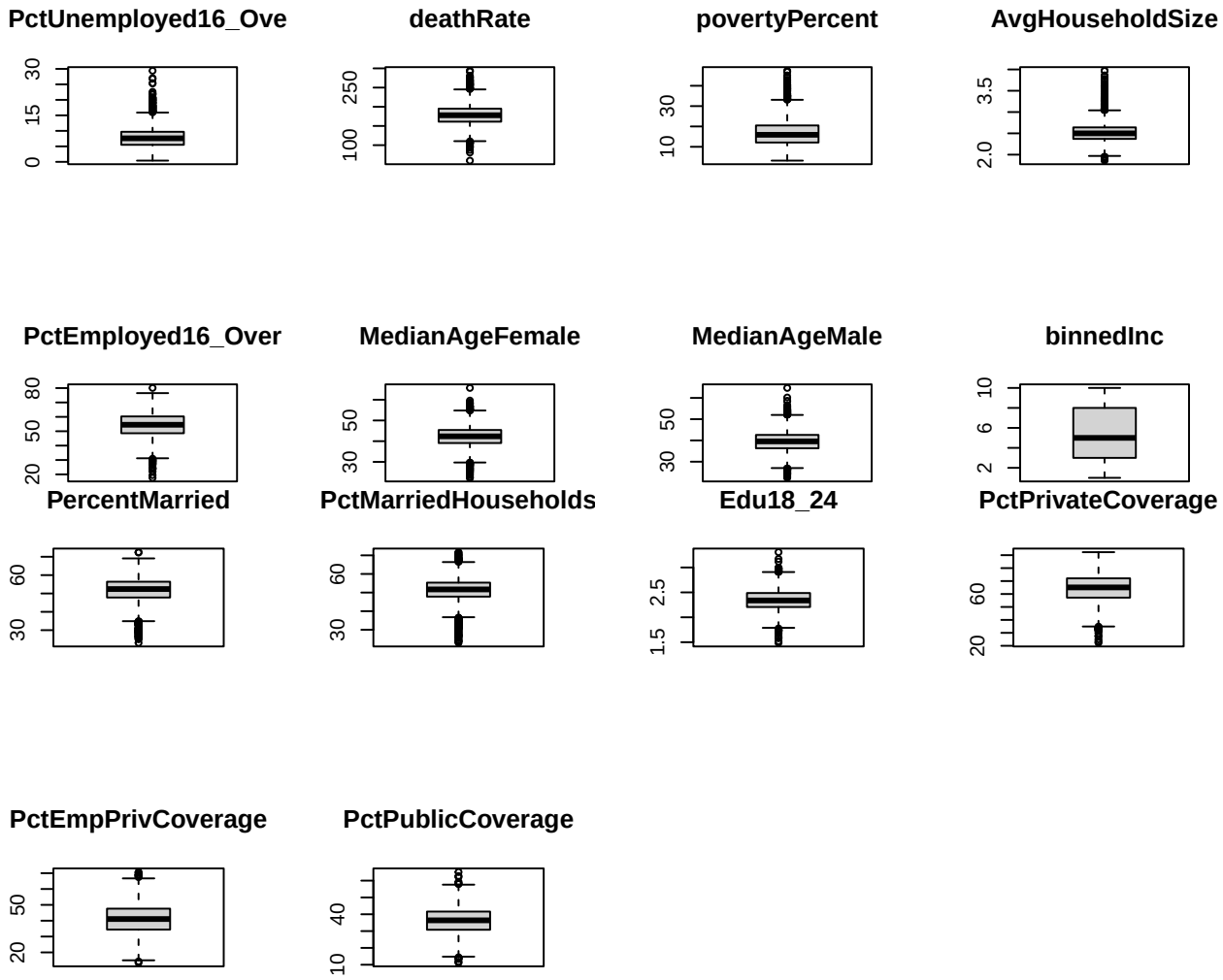


Figure 10: BoxPlots of our variables

3: Bivariate Plots

3.1: Plots Against Death Rate

```
# Checking linearity and heteroscedascity using scatter plots
par(mfrow = c(2,5))
with(cancer2, scatter.smooth(deathRate~PctEmployed16_Over, ylab = "deathRate",
                  xlab = "PctEmployed16_Over", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~PctUnemployed16_Over, ylab = "deathRate",
                  xlab = "PctUnemployed16_Over", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~PercentMarried, ylab = "deathRate",
                  xlab = "PercentMarried", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~PctMarriedHouseholds, ylab = "deathRate",
                  xlab = "PctMarriedHouseholds", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~Edu18_24, ylab = "deathRate",
                  xlab = "Edu18_24", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~binnedInc, ylab = "deathRate",
                  xlab = "binnedInc", lpars = list(col = "red", lwd = 2)))
```

```
with(cancer2, scatter.smooth(deathRate~povertyPercent, ylab = "deathRate",
                  xlab = "povertyPercent", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~PctPrivateCoverage, ylab = "deathRate",
                  xlab = "PctPrivateCoverage", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~PctPublicCoverage, ylab = "deathRate",
                  xlab = "PctPublicCoverage", lpars = list(col = "red", lwd = 2)))
with(cancer2, scatter.smooth(deathRate~PctEmpPrivCoverage, ylab = "deathRate",
                  xlab = "PctEmpPrivCoverage", lpars = list(col = "red", lwd = 2)))
```

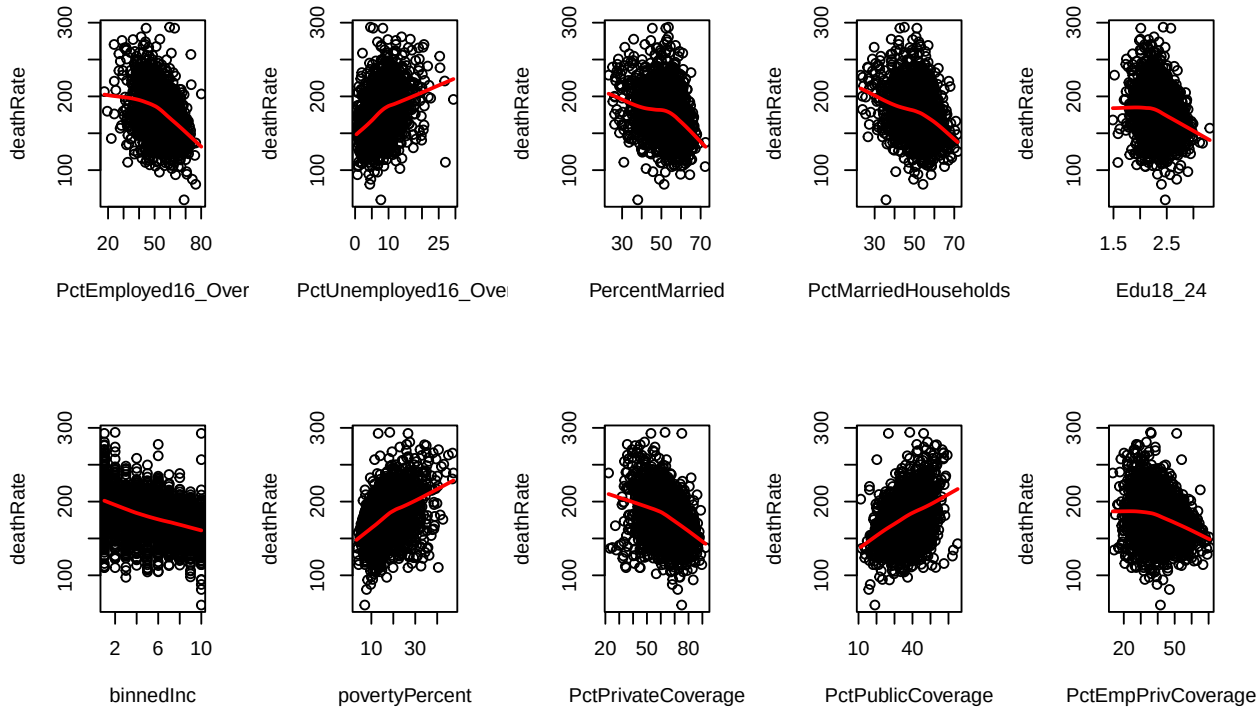


Figure 11: Plots showing deathRate against other variables

3.2: Multicollinearity

```
# To compare medIncome and binnedInc
with(cancer2, plot(binnedInc, medIncome,
main = "Comparison of medIncome and binnedInc", xlab = "binnedInc",
ylab = "medIncome"))
```

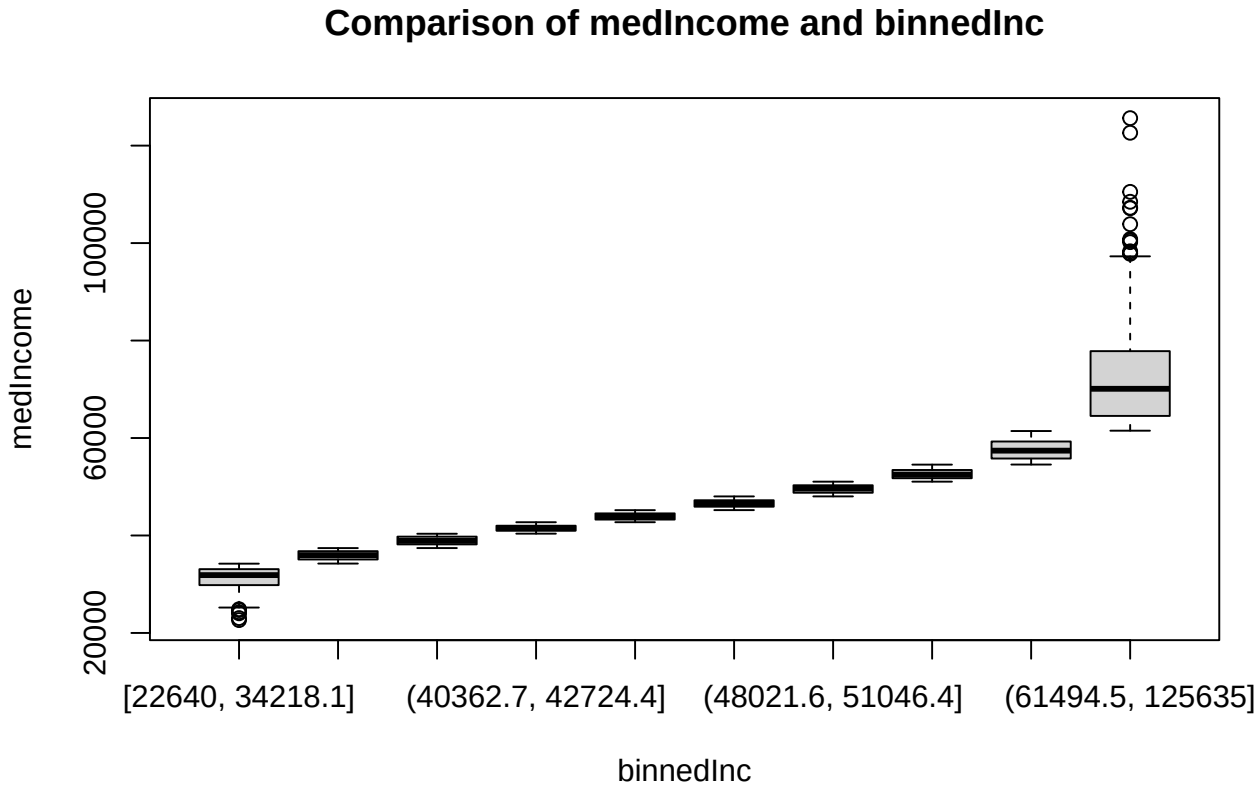


Figure 12: medIncome and binnedInc show very similar results

```
with(cancer2, cor.test(MedianAgeMale, MedianAgeFemale))
with(cancer2, cor.test(PercentMarried,PctMarriedHouseholds))
with(cancer2, cor.test(PctUnemployed16_Over, PctEmployed16_Over))
with(cancer2, cor.test(povertyPercent, PctEmployed16_Over))
with(cancer2, cor.test(povertyPercent, PctUnemployed16_Over))
with(cancer2, cor.test(povertyPercent, PctPrivateCoverage))
with(cancer2, cor.test(povertyPercent, PctEmpPrivCoverage))
with(cancer2, cor.test(povertyPercent, PctPublicCoverage))
```

```
##
##  Pearson's product-moment correlation
##
## data:  MedianAgeMale and 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
##
##
##  Pearson's product-moment correlation
##
## data:  PercentMarried and 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
##
```

```
##
##  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
##
##
##  Pearson's product-moment correlation
##
## data:  povertyPercent and 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:  povertyPercent and 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:  povertyPercent and 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:  povertyPercent and 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:  povertyPercent and 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
```

4: Multivariate Plots

4.1: Multicollinearity

```
# A correlation matrix plot highlighting strongly correlated variables(>0.6)
ggcorr(cancer2[,-c(1,4)], geom = "blank", label = TRUE, hjust = 0.9,
       label_size = 2.5, size = 2.5, label_round = 2) +
  geom_point(size = 10, aes(color = coefficient > 0,
                            alpha = abs(coefficient) > 0.6)) +
  scale_alpha_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +
  guides(color = "none", alpha = "none") +
  ggtitle("Correlation matrix of predictor variables")
```