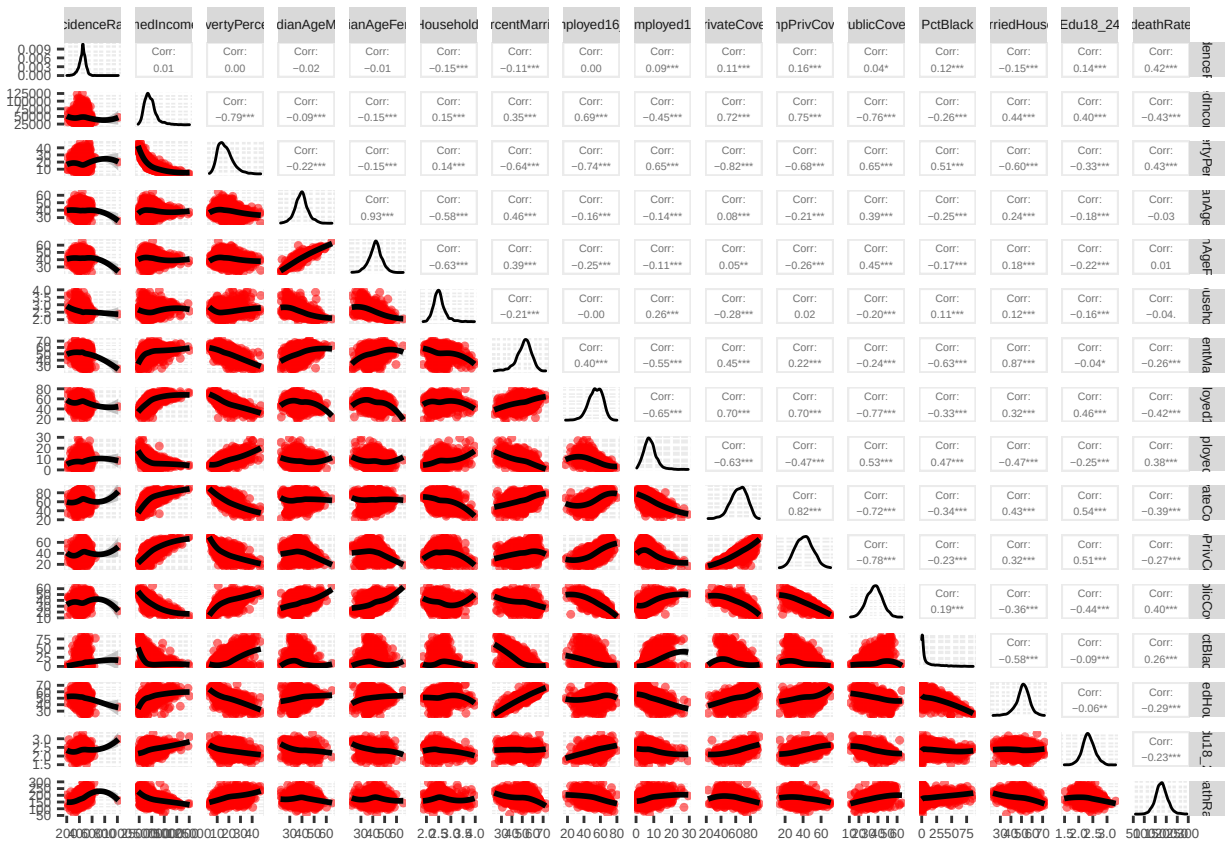
Stoopid Scatterplots

Alex

08/02/2022

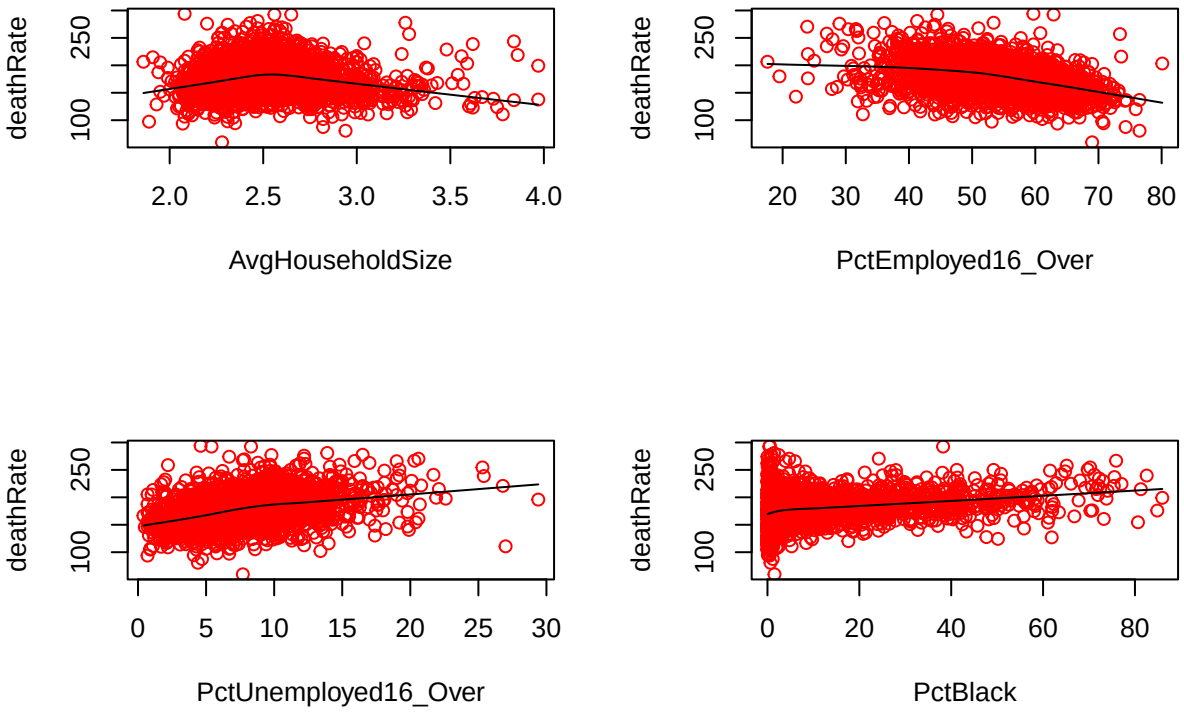
```
cancer1 <- cancer
cancer1$AvgHouseholdSize[cancer1$AvgHouseholdSize < 0.5] <- NA
cancer2 <- na.omit(cancer1)
```

```
mod_points=function(data,mapping,...) {
  ggally_smooth_loess(data, mapping,pch=20, ...) +
    theme(text = element_text(size=8))
}
mod_cor=function(data,mapping,...) {
  ggally_cor(data, mapping,size=1.5,align_percent=0.9, digits = 2) + scale_colour_manual(values = c("re
}
ggpairs(cancer2, columns = c(2:3, 5:18), mapping = aes(alpha = 0.6),
        upper = list(continuous=mod_cor),
        lower = list(continuous=wrap(mod_points, col = "red")),
        diag = list(continuous="densityDiag"),
        progress = FALSE) +
        theme(text=element_text(size=6))
```



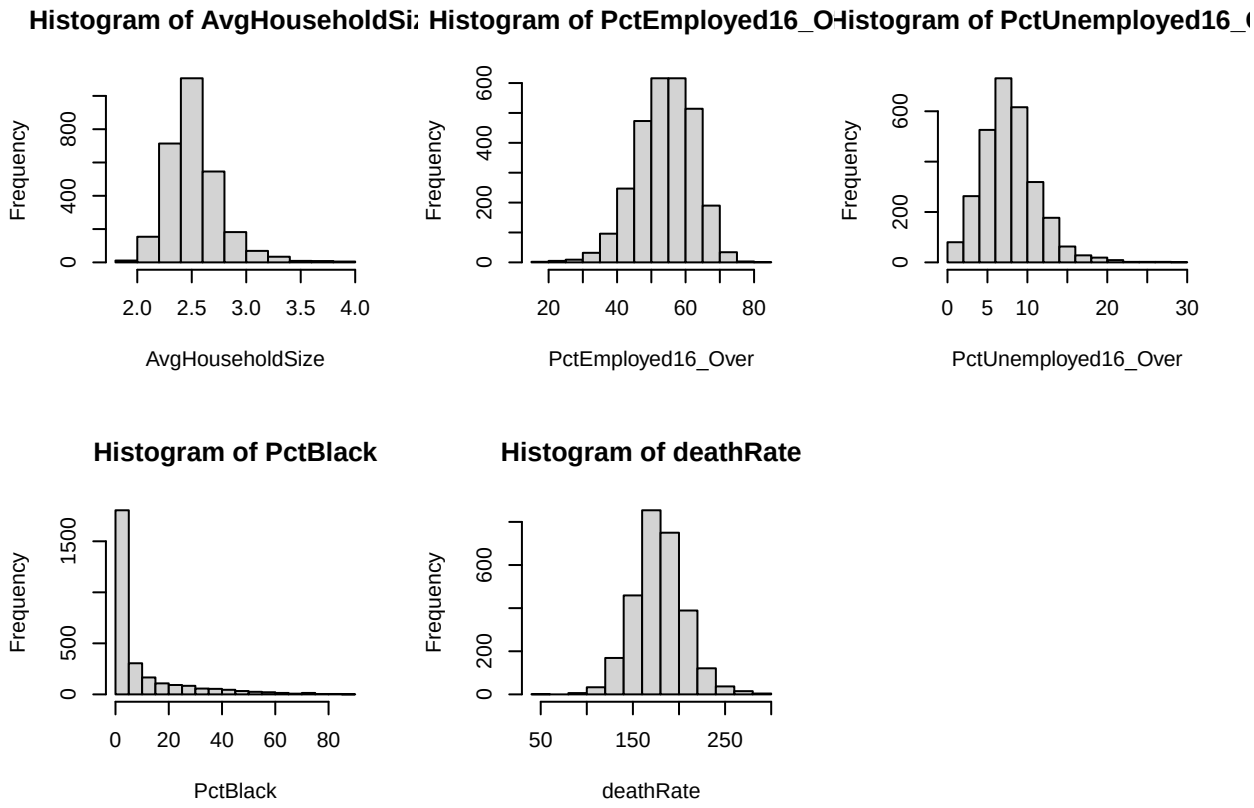
My Allocations

```
par(mfrow = c(2,2))
with(cancer2, scatter.smooth(deathRate~AvgHouseholdSize, col = "red"))
with(cancer2, scatter.smooth(deathRate~PctEmployed16_Over, col = "red"))
with(cancer2, scatter.smooth(deathRate~PctUnemployed16_Over, col = "red"))
with(cancer2, scatter.smooth(deathRate~PctBlack, col = "red"))
```

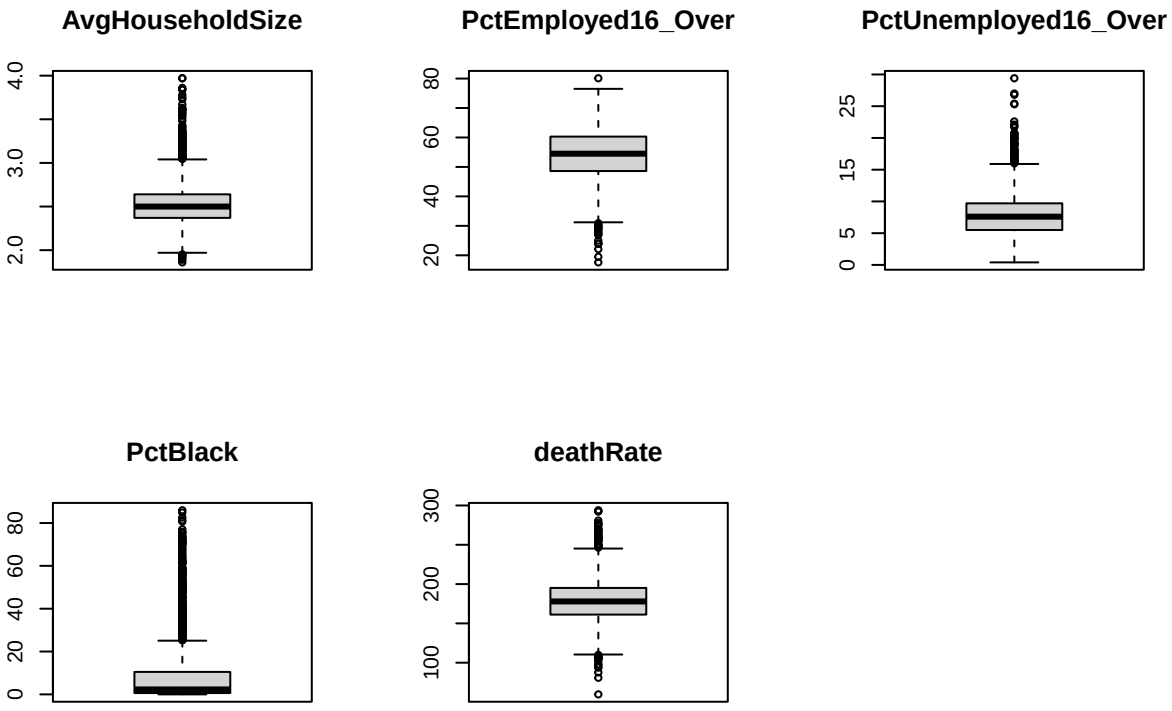


From the bivariate plots there is definite heteroskedasticity in pctBlack and for AvgHouseholdSize we see some non linearity. We see a concave shape so advising a more compex model, perhaps with a quadratic term might be advisable as the data is not monotonic.

```
par(mfrow = c(2,3))
with(cancer2, hist(AvgHouseholdSize))
with(cancer2, hist(PctEmployed16_Over))
with(cancer2, hist(PctUnemployed16_Over))
with(cancer2, hist(PctBlack))
with(cancer2, hist(deathRate))
```




```
par(mfrow = c(2,3))
with(cancer2, boxplot(AvgHouseholdSize, main = "AvgHouseholdSize"))
with(cancer2, boxplot(PctEmployed16_Over, main = "PctEmployed16_Over"))
with(cancer2, boxplot(PctUnemployed16_Over, main = "PctUnemployed16_Over"))
with(cancer2, boxplot(PctBlack, main = "PctBlack"))
with(cancer2, boxplot(deathRate, main = "deathRate"))
```



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.

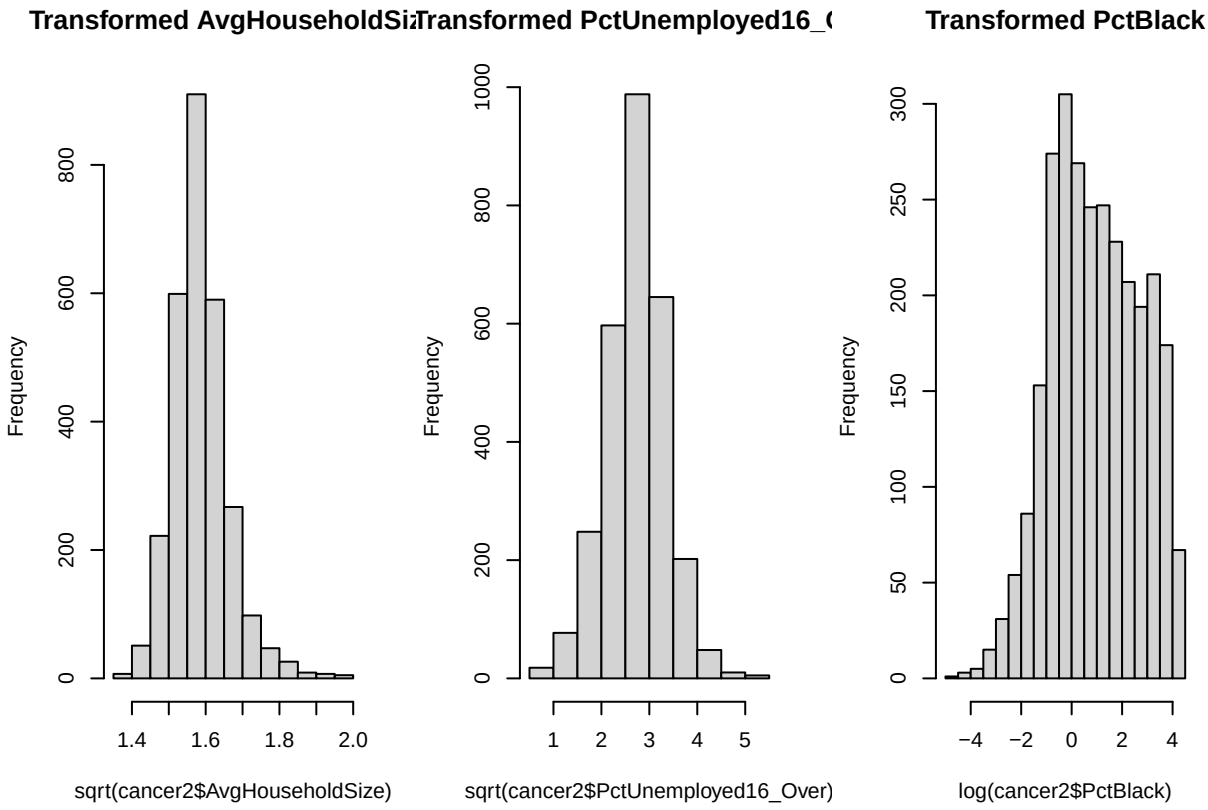
For heteroskedasticity we would need to perform further tests after fititng a model to check what kind of transformation we'd need to fix it.

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.

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



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

```
cancer4 <- cancer2
uwu2 <- str_split(cancer4$Geography, pattern = ", ")
cancer4$state <- rep(0, length(cancer4$Geography))
for(i in 1:length(cancer4$Geography)){
  cancer4$state[i] <- uwu2[[i]][2]
}

cancer4$county <- rep(0, length(cancer4$Geography))
for(i in 1:length(cancer4$Geography)){
  cancer4$county[i] <- uwu2[[i]][1]
}
cancer4$county[159] <- "Dona Ana County"
cancer4$county[775] <- "La Salle Parish"
```

```
cancer4$fips <- rep("0", length(cancer4$Geography))
for(i in 1:length(cancer4$Geography)){
  cancer4$fips[i] <- fips(cancer4$state[i], cancer4$county[i])
}
deathMap <- subset(cancer4, select = c("fips", "county", "deathRate"))
plot_usmap(data = deathMap,regions = "counties", values = "deathRate", include = cancer4$fips, color =
  scale_fill_continuous(low = "yellow", high = "red", name = "Death Rate", label = scales::comma) +
  labs(title = "Death rates in the United States") +
  theme(legend.position = "right")
```