Data Preparation

```
## Loading required package: carData

## Loading required package: rpart

## corrplot 0.94 loaded

## Warning: package 'PerformanceAnalytics' was built under R version 4.4.2

## Loading required package: xts

## Warning: package 'xts' was built under R version 4.4.2

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend

##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind

##
## ######################### Warning from 'xts' package ##########################
## #                                                                             #
## # The dplyr lag() function breaks how base R's lag() function is supposed to  #
## # work, which breaks lag(my_xts). Calls to lag(my_xts) that you type or       #
## # source() into this session won't work correctly.                            #
## #                                                                             #
## # Use stats::lag() to make sure you're not using dplyr::lag(), or you can add #
## # conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop           #
```

```
## # dplyr from breaking base R's lag() function.                              #
## #                                                                            #
## # Code in packages is not affected. It's protected by R's namespace mechanism #
## # Set `options(xts.warn_dplyr_breaks_lag = FALSE)` to suppress this warning.  #
## #                                                                            #
## ###############################################################################

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:xts':
##
##     first, last

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

First we import the data and save it as the variable "df" for future modifications.

```r
par(mfrow=c(1,1))
df <- read.csv("data/train.csv")
```

Variable analysis

We perform descriptive analysis for each variable of this data, a data quality report , profiling and imputation if needed.

```r
colnames(df)
```

```
##  [1] "avganncount"             "avgdeathsperyear"
##  [3] "target_deathrate"        "incidencerate"
##  [5] "medincome"               "popest2015"
##  [7] "povertypercent"          "studypercap"
##  [9] "binnedinc"               "medianage"
## [11] "medianagemale"           "medianagefemale"
## [13] "geography"               "percentmarried"
## [15] "pctnohs18_24"            "pcths18_24"
## [17] "pctsomecol18_24"         "pctbachdeg18_24"
## [19] "pcths25_over"            "pctbachdeg25_over"
## [21] "pctemployed16_over"      "pctunemployed16_over"
## [23] "pctprivatecoverage"      "pctprivatecoveragealone"
## [25] "pctempprivcoverage"      "pctpubliccoverage"
```

```
## [27] "pctpubliccoveragealone"  "pctwhite"
## [29] "pctblack"                "pctasian"
## [31] "pctotherrace"            "pctmarriedhouseholds"
## [33] "birthrate"
```

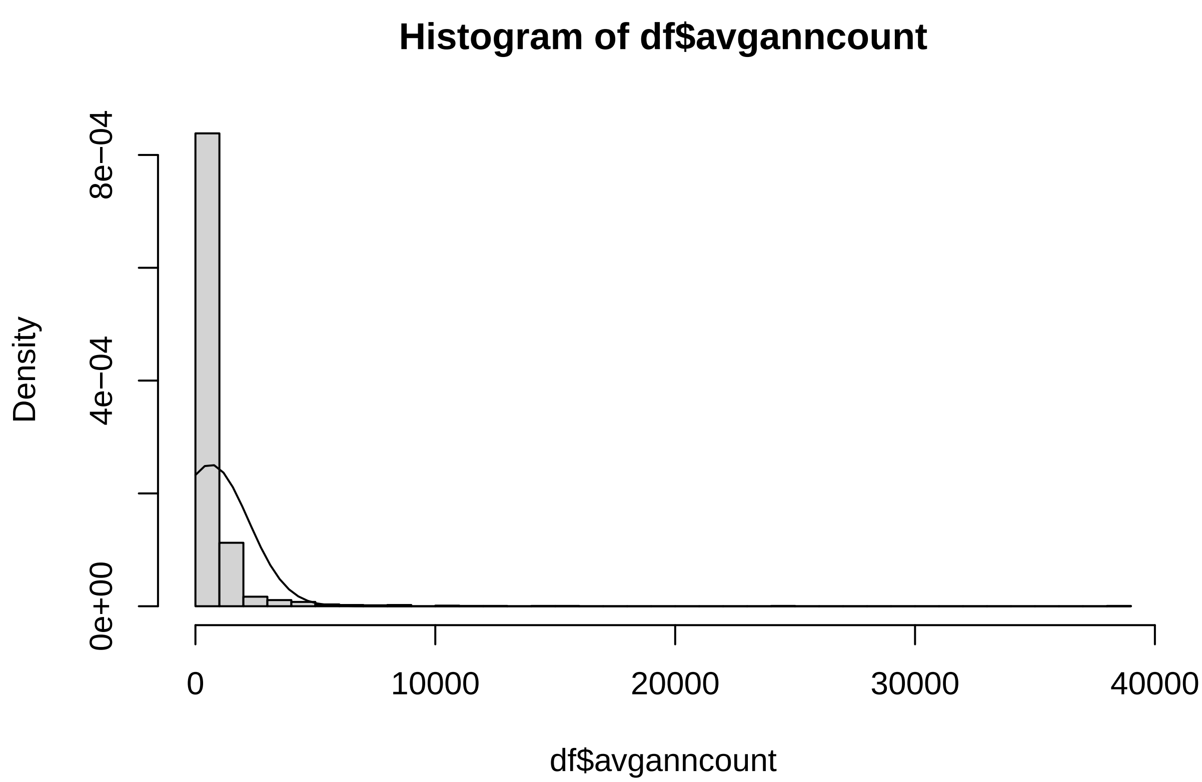
Variable 1 - avganncount

This is a continuous ratio variable. The data does not look normally distributed, which is confirmed by the near-null p-value of the shapiro normallity test. A histogram is used to visualize the data. The variable contains no missing values thus imputation is not needed. It contains 273 outliers (out of which 252 severe), all on the higher end of the spectrum. We create an additional ordinal factor "f.avganncount" to create a discretisation according to the quartiles.

```r
summary(df$avganncount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.0    80.0   175.0   623.2   509.0 38150.0
```

```r
hist(df$avganncount, breaks = 30, freq = F)
curve(dnorm(x, mean(df$avganncount), sd(df$avganncount)), add = T)
```



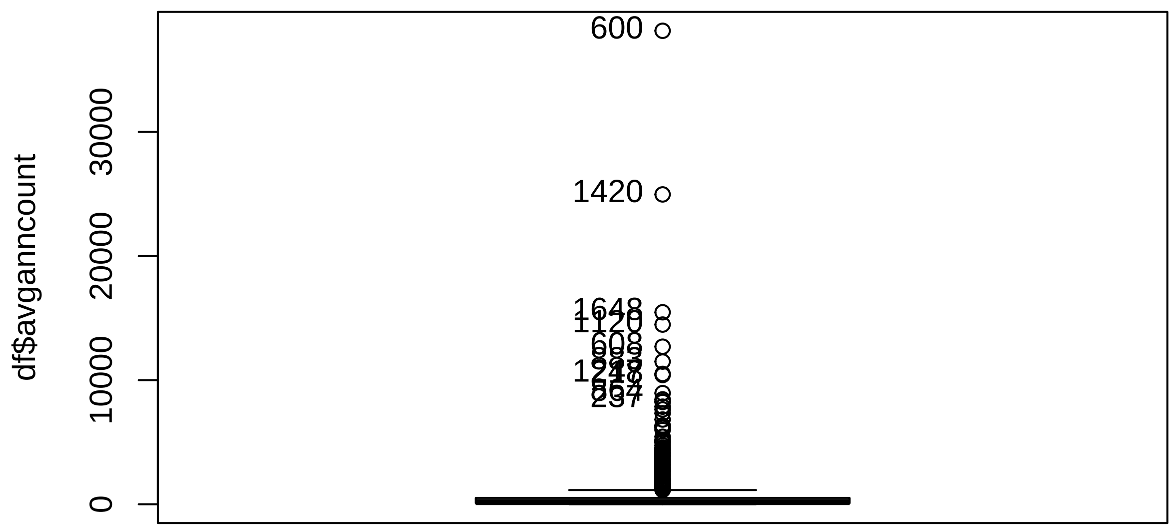

```r
shapiro.test(df$avganncount)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$avganncount
## W = 0.33377, p-value < 2.2e-16
```

```r
sum(is.na(df$avganncount))
```

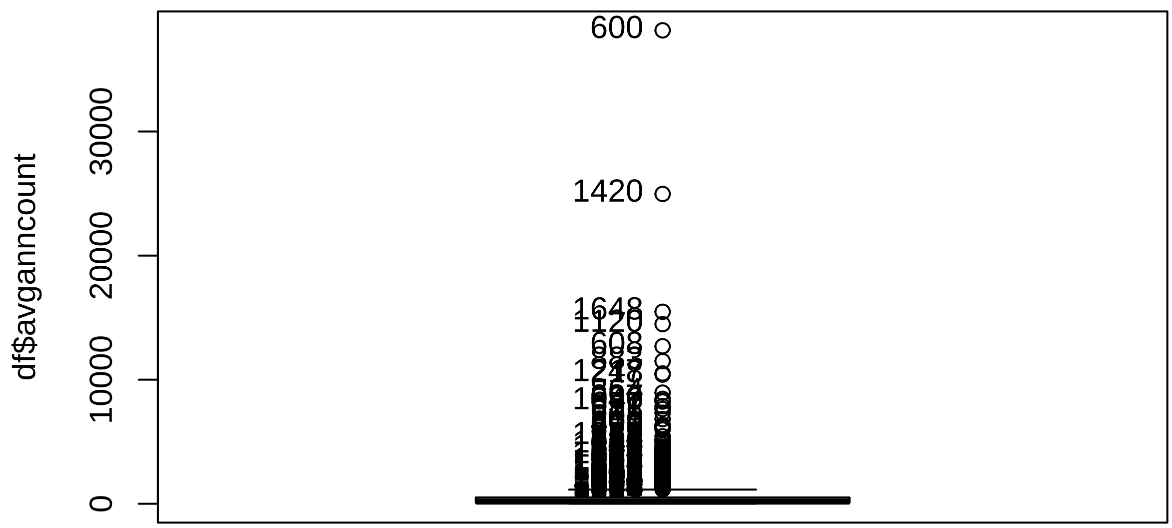
```
## [1] 0
```

```r
Boxplot(df$avganncount)
```




```
##  [1]  600 1420 1648 1120  608  883 1247  218  864  237
```

```
length(Boxplot(df$avganncount, id = list(n=Inf)))
```



```
## [1] 273
```

```
sevout_avganncount = (quantile(df$avganncount,0.25)+(3*((quantile(df$avganncount,0.75)-quantile(df$avgar
length(which(df$avganncount > sevout_avganncount))
```

```
## [1] 252
```

```
df$f.avganncount <- ifelse(df$avganncount <= 80.0, 1, ifelse(df$avganncount > 80.0 & df$avganncount <= 1
df$f.avganncount <- factor(df$f.avganncount, labels=c("LowCaseCount","LowMidCaseCount","HighMidCaseCount
table(df$f.avganncount)
```

```
##
##     LowCaseCount  LowMidCaseCount HighMidCaseCount    HighCaseCount
##              460              458              455              458
```

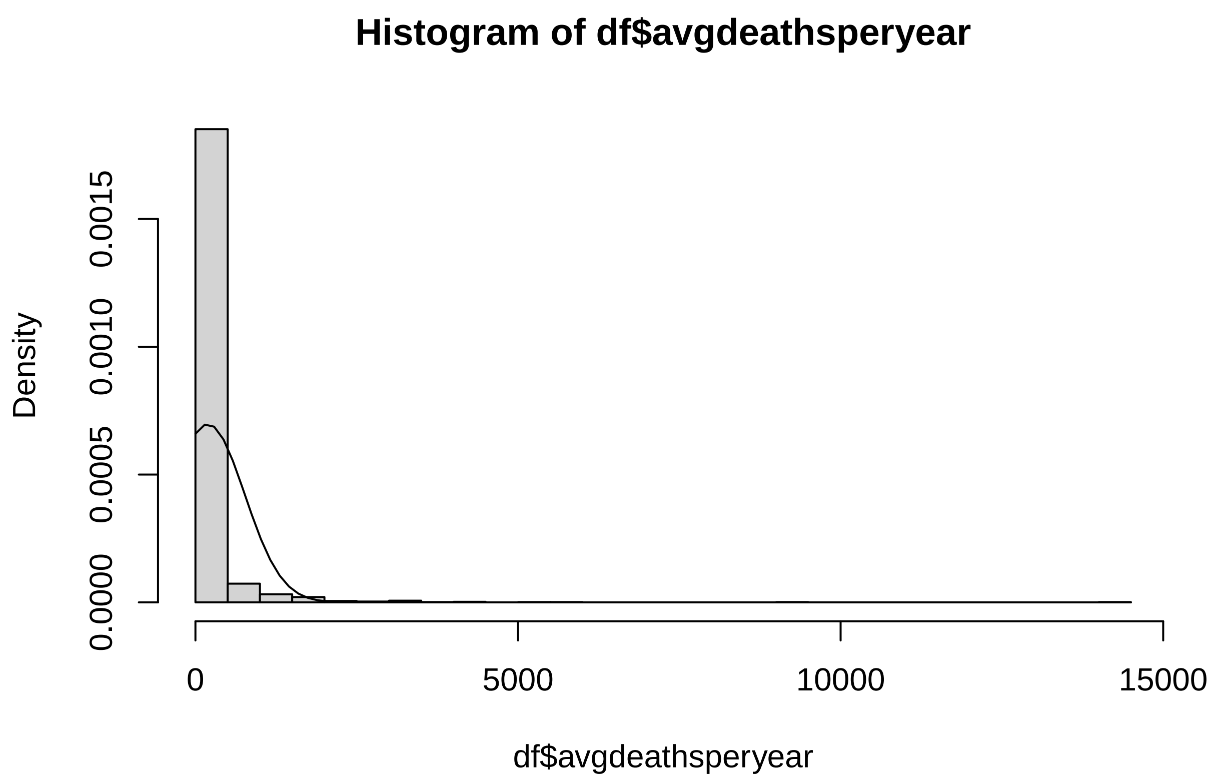
Variable 2 - avgdeathsperyear

This is also a continuous ratio variable similar to variable 1. The data does not look normally distributed, which is confirmed by the near-null p-value of the shapiro normallity test. Again a histogram is used to visualize the data. The variable contains no missing values thus imputation is not needed. It contains 225 outliers (out of which 178 severe), all on the higher end of the spectrum. We create an additional ordinal factor "f.avgdeathsperyear" to create a discretisation according to the quartiles.

```r
summary(df$avgdeathsperyear)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    29.0    62.0   191.6   140.5 14010.0
```

```r
hist(df$avgdeathsperyear, breaks = 30, freq = F)
curve(dnorm(x, mean(df$avgdeathsperyear), sd(df$avgdeathsperyear)), add = T)
```



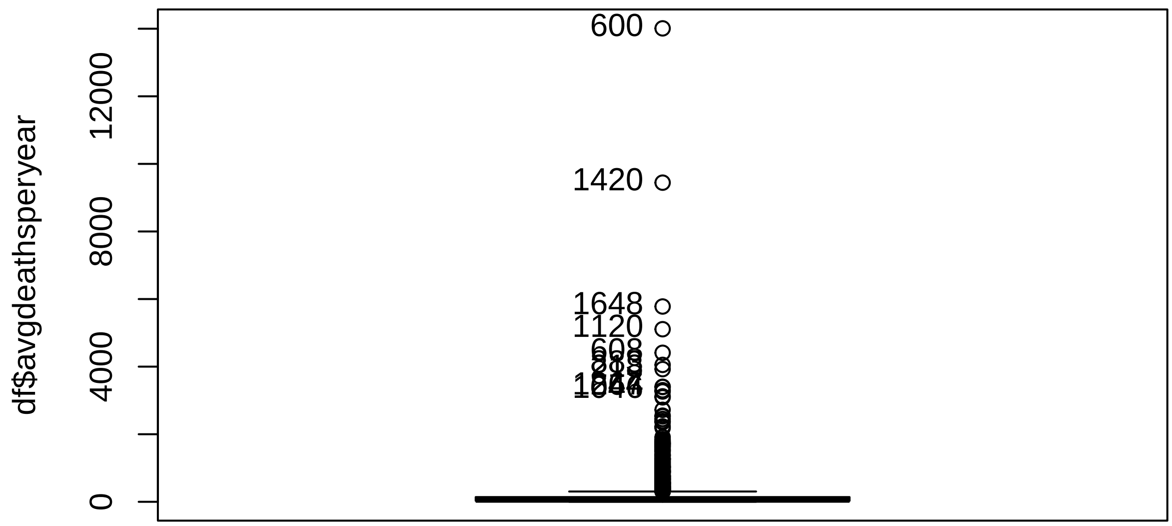

```r
shapiro.test(df$avgdeathsperyear)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$avgdeathsperyear
## W = 0.26769, p-value < 2.2e-16
```

```r
sum(is.na(df$avgdeathsperyear))
```

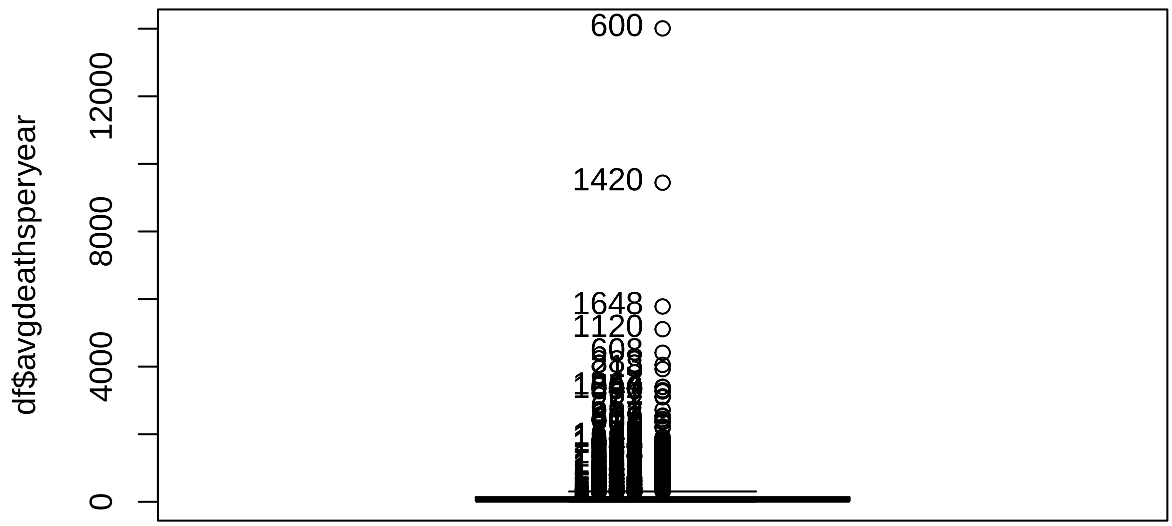
```
## [1] 0
```

```r
Boxplot(df$avgdeathsperyear)
```




```
## [1]  600 1420 1648 1120  608  883  218 1247  864 1046
```

```r
length(Boxplot(df$avgdeathsperyear, id = list(n=Inf)))
```




```
## [1] 225
```

```
sevout_avgdeathsperyear = (quantile(df$avgdeathsperyear,0.25)+(3*((quantile(df$avgdeathsperyear,0.75)-qu
length(which(df$avgdeathsperyear > sevout_avgdeathsperyear))
```

```
## [1] 178
```

```
df$f.avgdeathsperyear <- ifelse(df$avgdeathsperyear <= 29.0, 1, ifelse(df$avgdeathsperyear > 29.0 & df$a
df$f.avgdeathsperyear <- factor(df$f.avgdeathsperyear, labels=c("LowMortCount","LowMidMortCount","HighMi
table(df$f.avgdeathsperyear)
```

```
##
##     LowMortCount  LowMidMortCount HighMidMortCount    HighMortCount
##              462              455              456              458
```

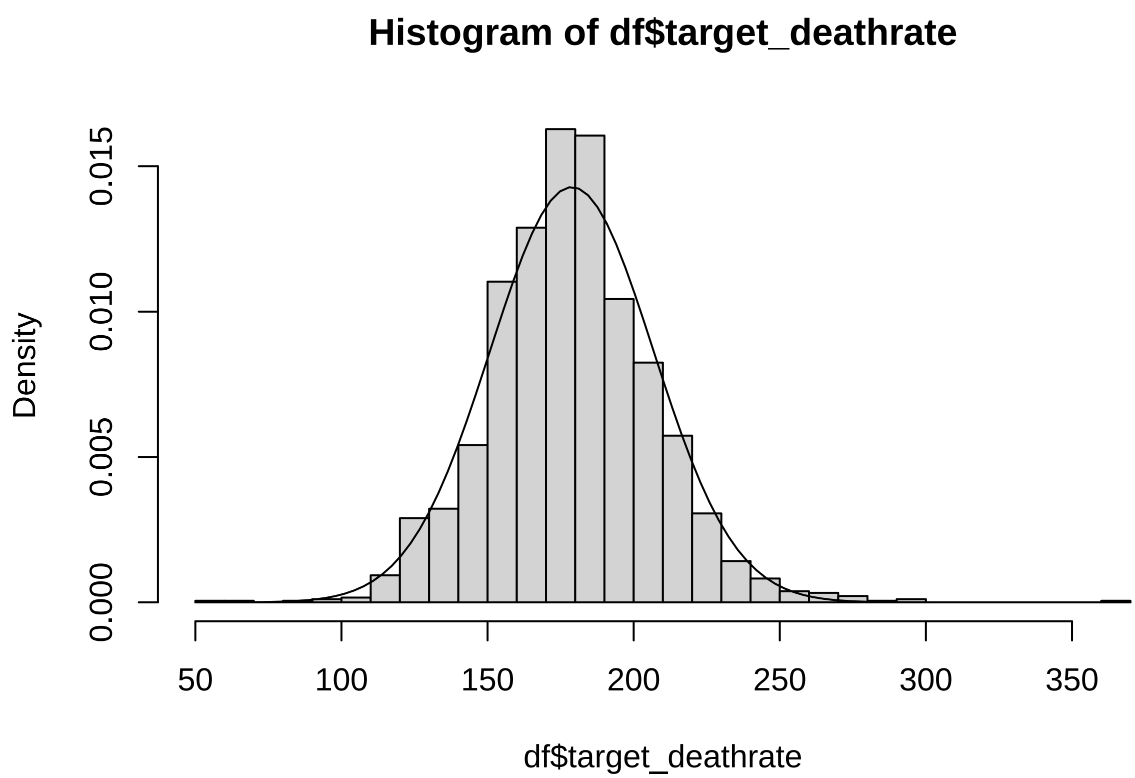
Variable 3 - target_deathrate

This is the response variable. This is also a continuous ratio variable similar to the previous variables. The data looks normally distributed, but it is not and will be further discussed in the next section. It contains no missing values thus imputation is not needed. It contains 35 outliers (out of which 11 severe). We create an additional ordinal factor “f.deathrate” to create a discretisation according to the quartiles.

```r
summary(df$target_deathrate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    59.7   161.3   178.3   178.8   195.3   362.8
```

```r
hist(df$target_deathrate, breaks = 30, freq = F)
curve(dnorm(x, mean(df$target_deathrate), sd(df$target_deathrate)), add = T)
```



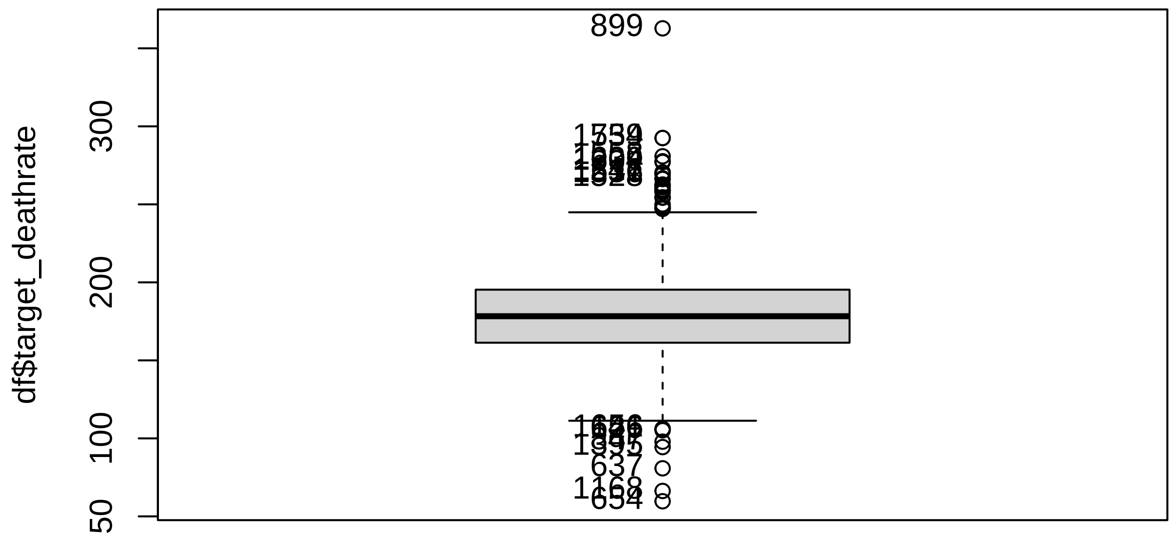

```r
shapiro.test(df$target_deathrate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$target_deathrate
## W = 0.98647, p-value = 4.149e-12
```

```r
sum(is.na(df$target_deathrate))
```

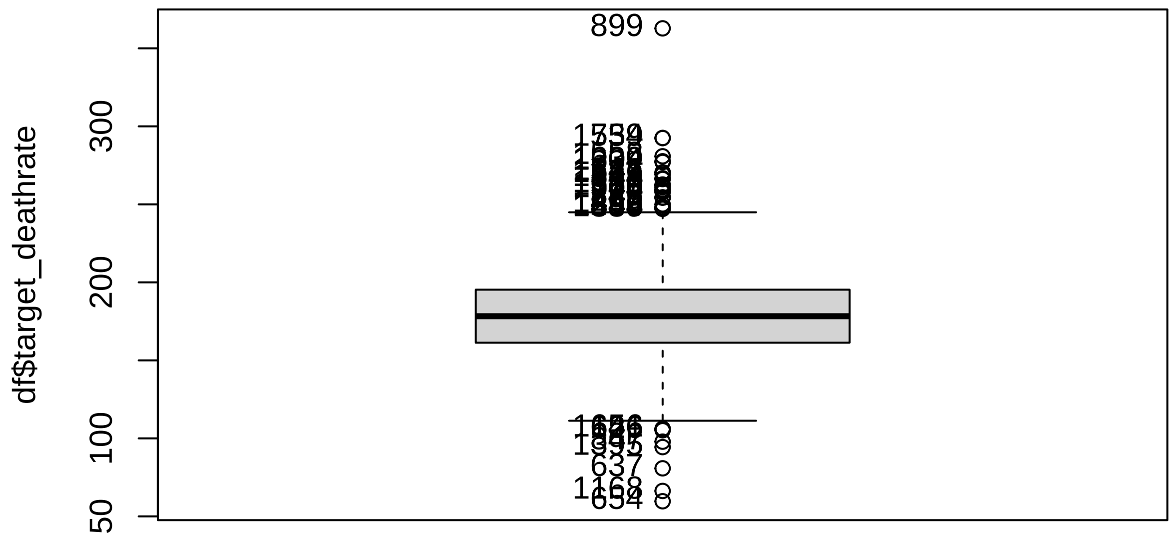
```
## [1] 0
```

```r
Boxplot(df$target_deathrate)
```




```
##  [1]  626  637  651  654  847 1146 1168 1395  899  734 1559 1558 1639 1304 1211
## [16] 1547 1536 1528
```

```r
length(Boxplot(df$target_deathrate, id = list(n=Inf)))
```




```
## [1] 35
```

```
sevout_deathrate = (quantile(df$target_deathrate,0.25)+(3*((quantile(df$target_deathrate,0.75)-quantile
length(which(df$target_deathrate > sevout_deathrate))
```

```
## [1] 11
```

```
df$f.deathrate <- ifelse(df$target_deathrate <= 161.3, 1, ifelse(df$target_deathrate > 161.3 & df$targe
df$f.deathrate <- factor(df$f.deathrate, labels=c("LowDeathrate","LowMidDeathrate","HighMidDeathrate","
table(df$f.deathrate)
```

```
##
##     LowDeathrate  LowMidDeathrate HighMidDeathrate    HighDeathrate
##              459              459              456              457
```

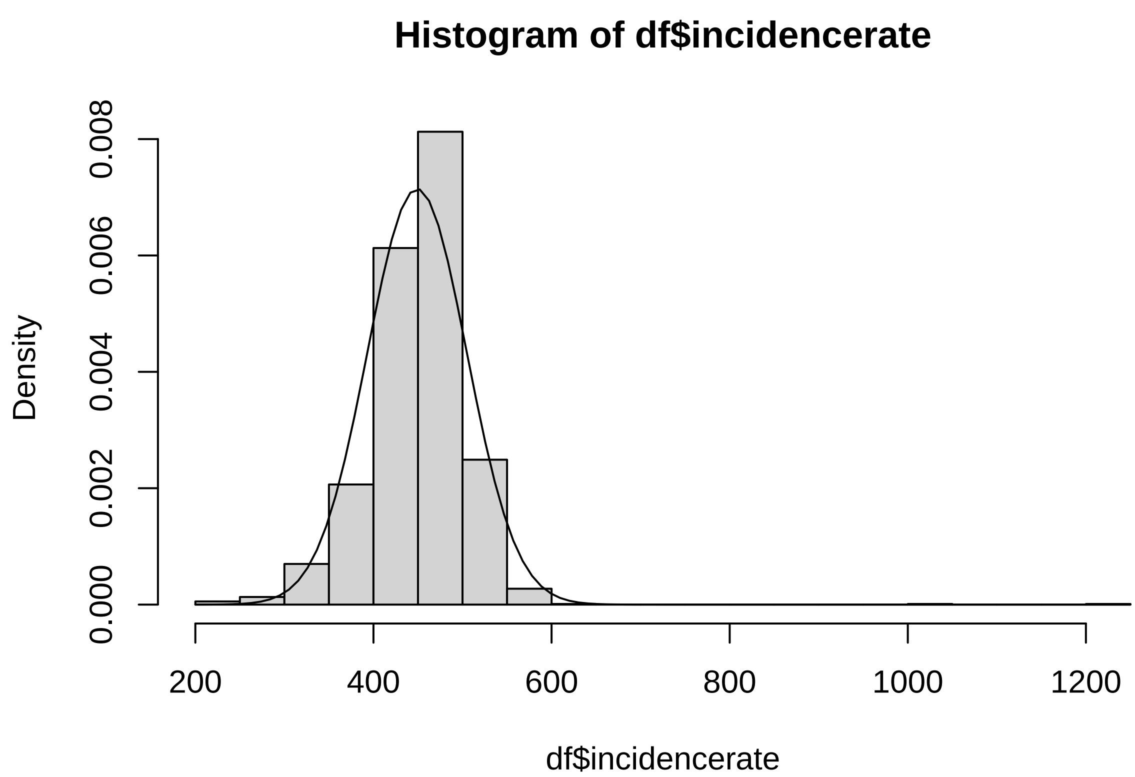
Variable 4 - incidencerate

We have another continuous ratio variable similar to the previous variables. It is not normally distributed according to the Shapiro test. It contains no missing values thus imputation is not needed. It contains 60 outliers (out of which 3 severe) in both the higher and the lower ends of the spectrum. We create an additional ordinal factor "f.incidencerate".

```r
summary(df$incidencerate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   201.3   421.4   453.5   449.0   481.3  1206.9
```

```r
hist(df$incidencerate, breaks = 30, freq = F)
curve(dnorm(x, mean(df$incidencerate), sd(df$incidencerate)), add = T)
```



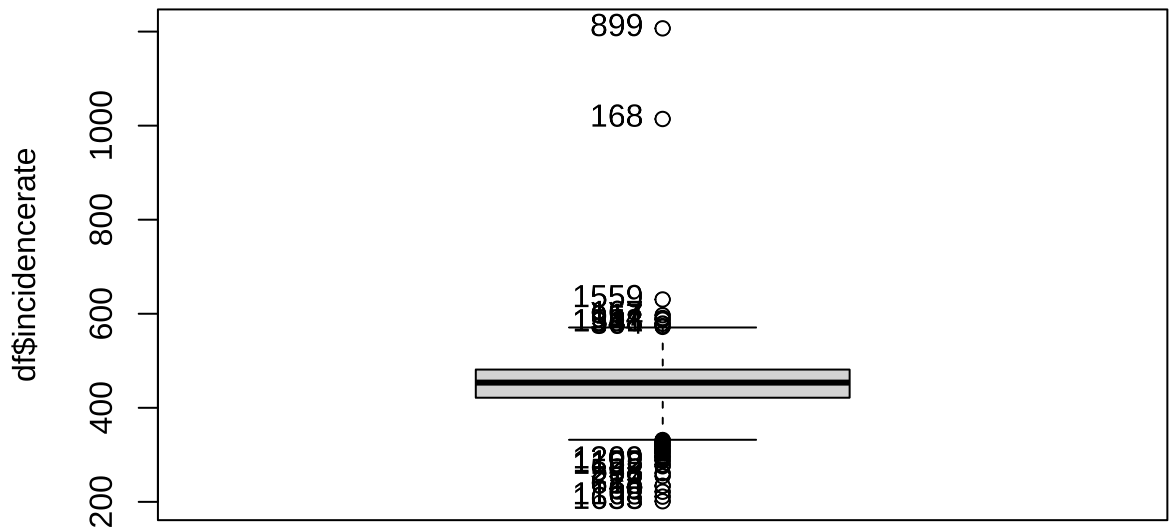

```r
shapiro.test(df$incidencerate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$incidencerate
## W = 0.89577, p-value < 2.2e-16
```

```r
sum(is.na(df$incidencerate))
```

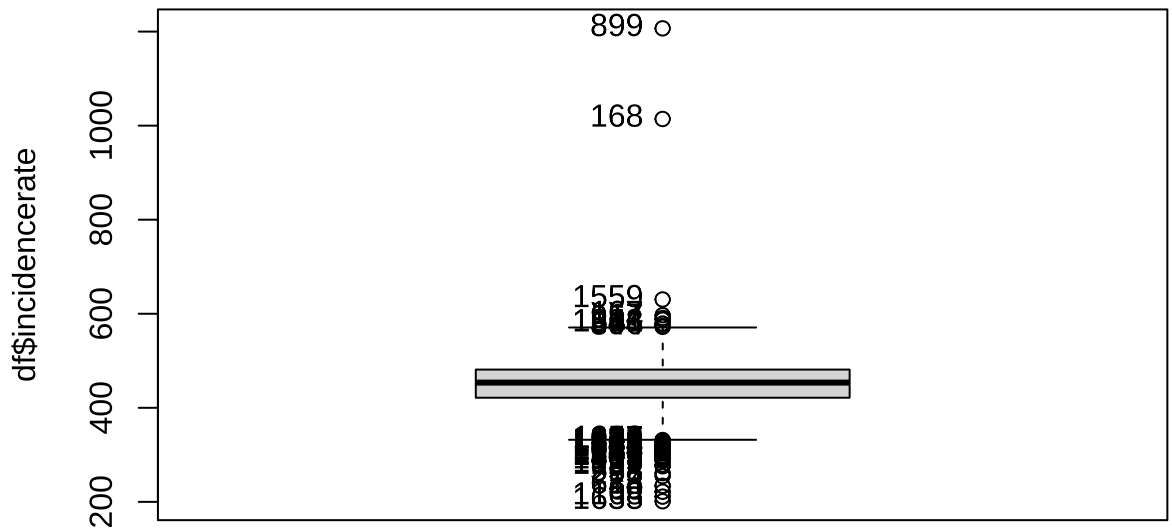
```
## [1] 0
```

```r
Boxplot(df$incidencerate)
```




```
##  [1] 1633 1168   60   18  634  295  558 1122 1155 1209  899  168 1559  167   17
## [16]  954 1558 1548 1541  364
```

```
length(Boxplot(df$incidencerate, id = list(n=Inf)))
```




```
## [1] 60
```

```r
sevout_incidencerate = (quantile(df$incidencerate,0.25)+(3*((quantile(df$incidencerate,0.75)-quantile(d
length(which(df$incidencerate > sevout_incidencerate))
```

```
## [1] 3
```

```r
df$f.incidencerate <- ifelse(df$incidencerate <= 421.4, 1, ifelse(df$incidencerate > 421.4 & df$incidenc
df$f.incidencerate <- factor(df$f.incidencerate, labels=c("LowDiagnPerCap","LowMidDiagnPerCap","HighMid
table(df$f.incidencerate)
```

```
##
##     LowDiagnPerCap  LowMidDiagnPerCap HighMidDiagnPerCap    HighDiagnPerCap
##                460                409                504                458
```

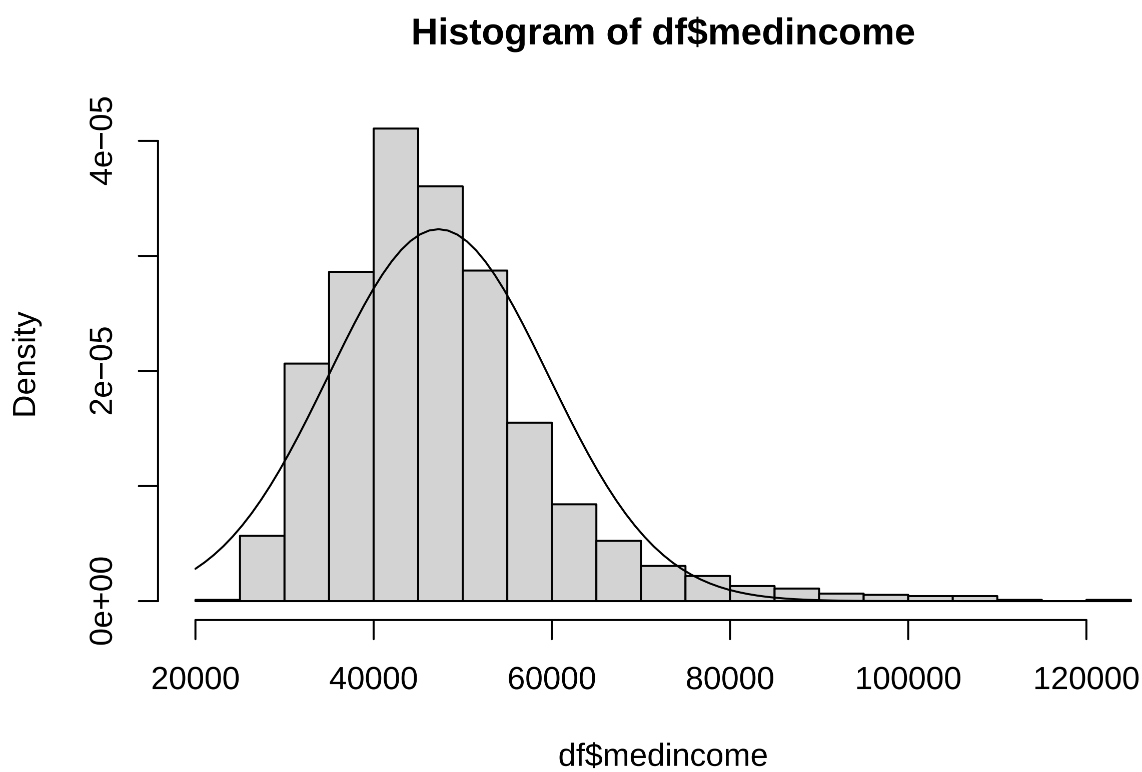
Variable 5 - medincome

Very similar to all the previous variables we have a continuous ratio variable not normally distributed with 0 missing values, 69 outliers (44 of them severe), all on the higher end. We create an additional ordinal factor "f.medincome".

```r
summary(df$medincome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22640   39031   45454   47278   52612  122641
```

```r
hist(df$medincome, breaks = 30, freq = F)
curve(dnorm(x, mean(df$medincome), sd(df$medincome)), add = T)
```



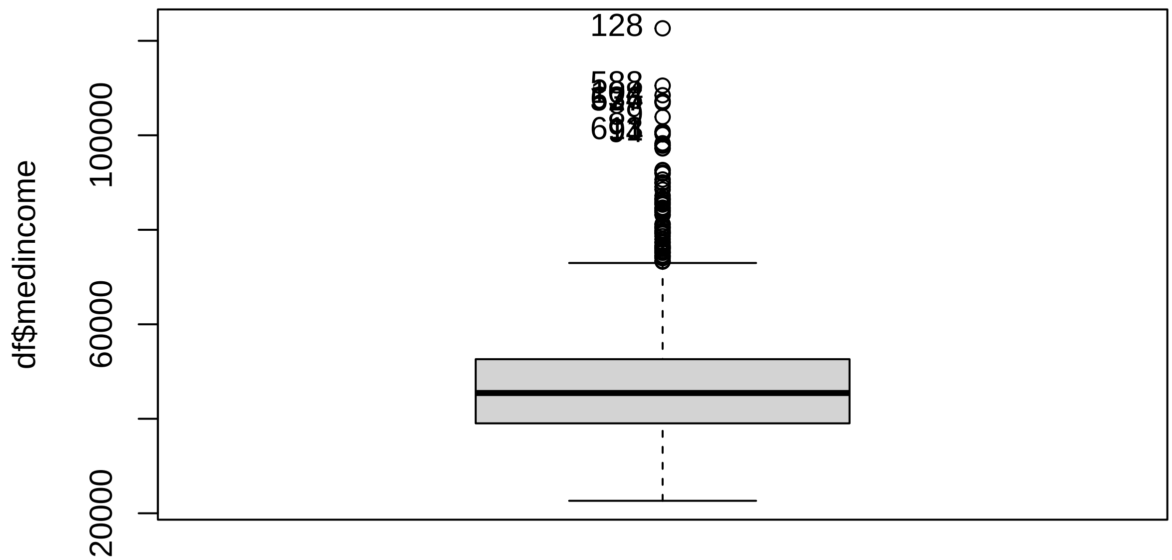

```r
shapiro.test(df$medincome)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$medincome
## W = 0.9105, p-value < 2.2e-16
```

```r
sum(is.na(df$medincome))
```

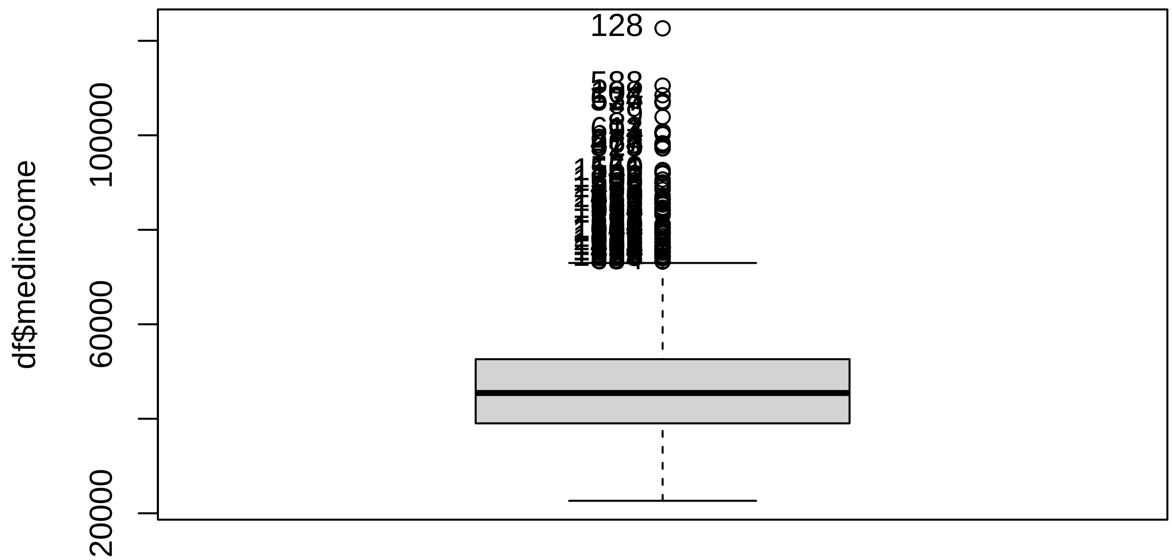
```
## [1] 0
```

```r
Boxplot(df$medincome)
```




```
##  [1] 128 588 104 636 574 527  89 613  91  94
```

```r
length(Boxplot(df$medincome, id = list(n=Inf)))
```




```
## [1] 69
```

```r
sevout_medincome = (quantile(df$medincome,0.25)+(3*((quantile(df$medincome,0.75)-quantile(df$medincome,0
length(which(df$medincome > sevout_medincome))
```

```
## [1] 44
```

```r
df$f.medincome <- ifelse(df$medincome <= 39031, 1, ifelse(df$medincome > 39031 & df$medincome <= 45454,
df$f.medincome <- factor(df$f.medincome, labels=c("LowMedianInc","LowMidMedianInc","HighMidMedianInc","H
table(df$f.medincome)
```

```
##
##     LowMedianInc  LowMidMedianInc HighMidMedianInc    HighMedianInc
##              458              458              457              458
```

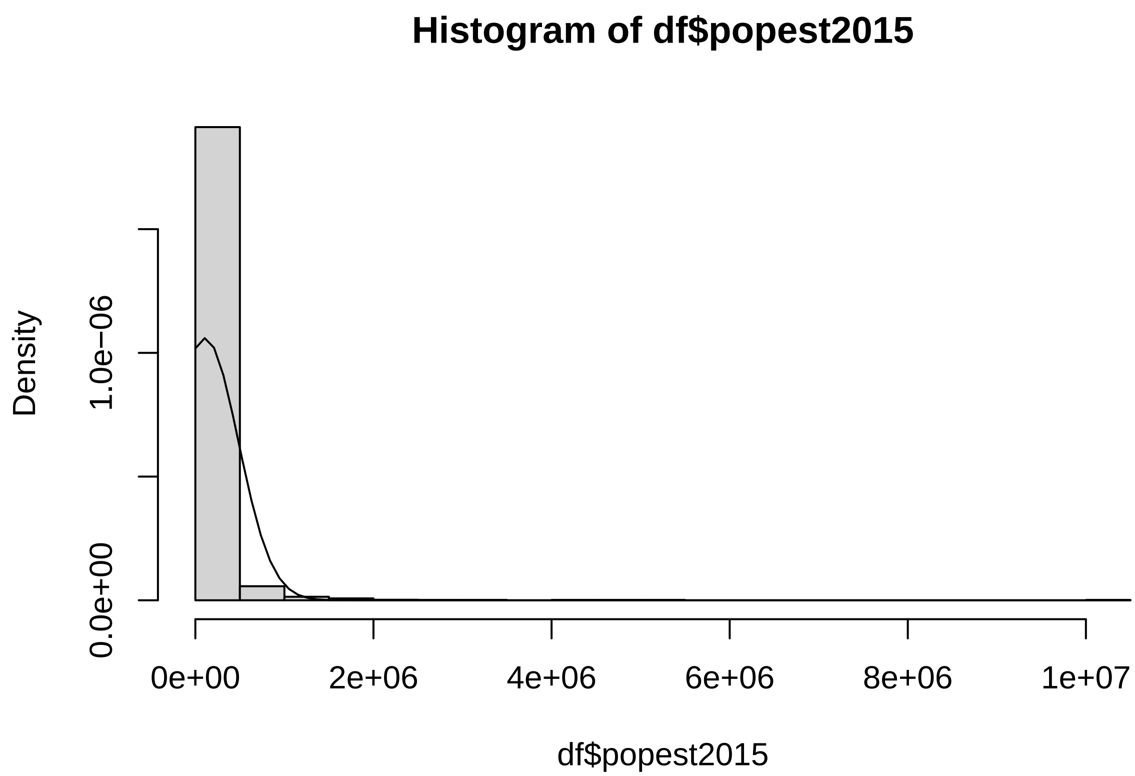
Variable 6 - popest2015

Another continuous ratio variable not normally distributed with 0 missing values, 252 outliers (210 of them severe), all on the higher end. We create an additional ordinal factor “f.popest2015”.

```r
summary(df$popest2015)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      829    12191    27158   106841    66880 10170292
```

```r
hist(df$popest2015, breaks = 30, freq = F)
curve(dnorm(x, mean(df$popest2015), sd(df$popest2015)), add = T)
```



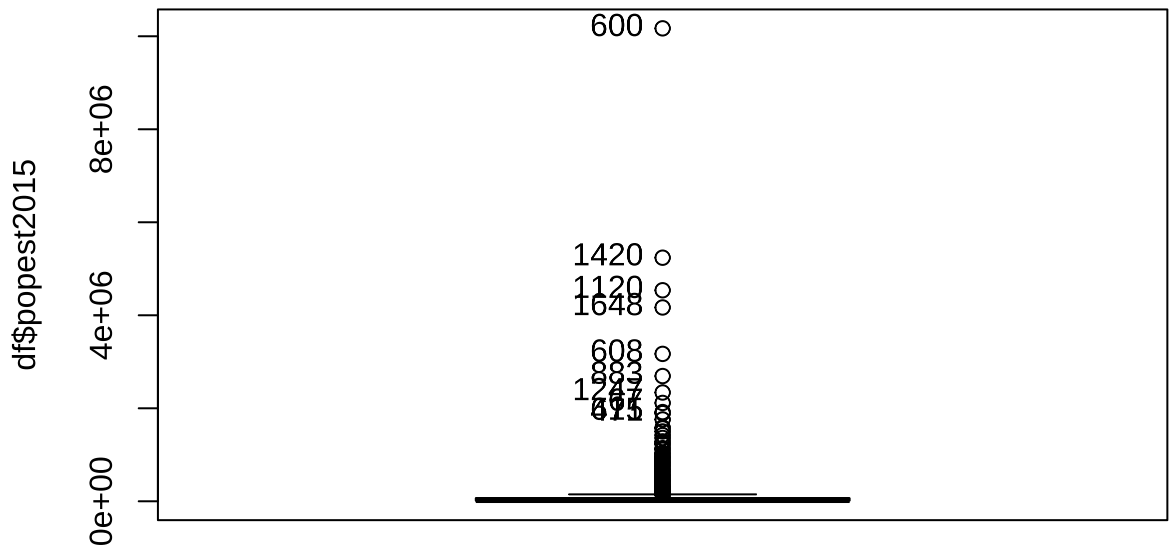

```r
shapiro.test(df$popest2015)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$popest2015
## W = 0.22666, p-value < 2.2e-16
```

```r
sum(is.na(df$popest2015))
```

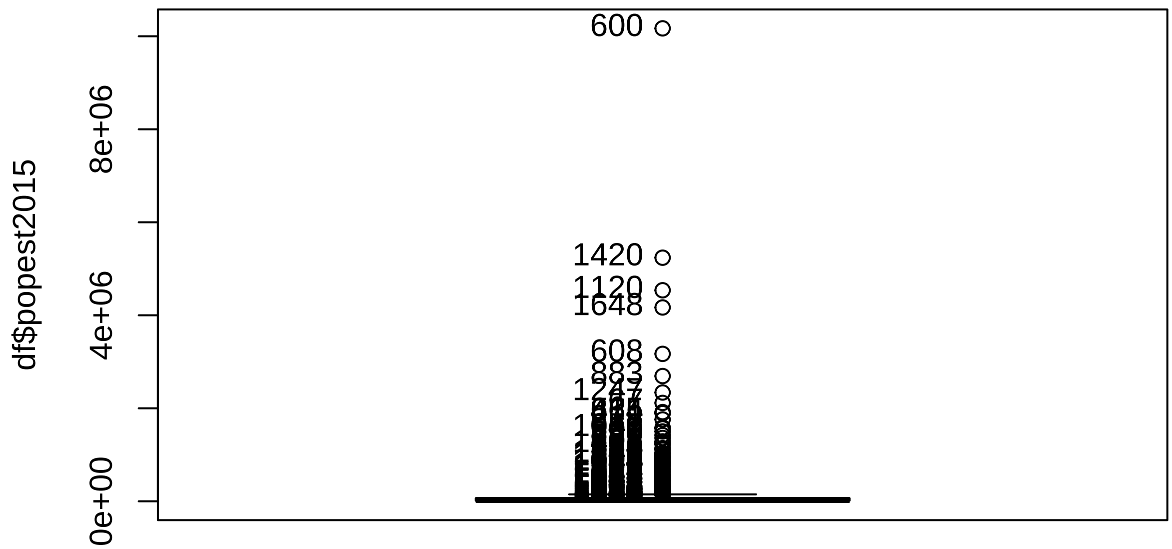
```
## [1] 0
```

```r
Boxplot(df$popest2015)
```




```
## [1]  600 1420 1120 1648  608  883 1247   67  615  471
```

```r
length(Boxplot(df$popest2015, id = list(n=Inf)))
```




```
## [1] 252
```

```
sevout_popest2015 = (quantile(df$popest2015,0.25)+(3*((quantile(df$popest2015,0.75)-quantile(df$popest20
length(which(df$popest2015 > sevout_popest2015))
```

```
## [1] 210
```

```
df$f.popest2015 <- ifelse(df$popest2015 <= 12191, 1, ifelse(df$popest2015 > 12191 & df$popest2015 <= 27
df$f.popest2015 <- factor(df$f.popest2015, labels=c("LowPop","LowMidPop","HighMidPop","HighPop"), order
table(df$f.popest2015)
```

```
##
##     LowPop  LowMidPop HighMidPop    HighPop
##        458        458        457        458
```

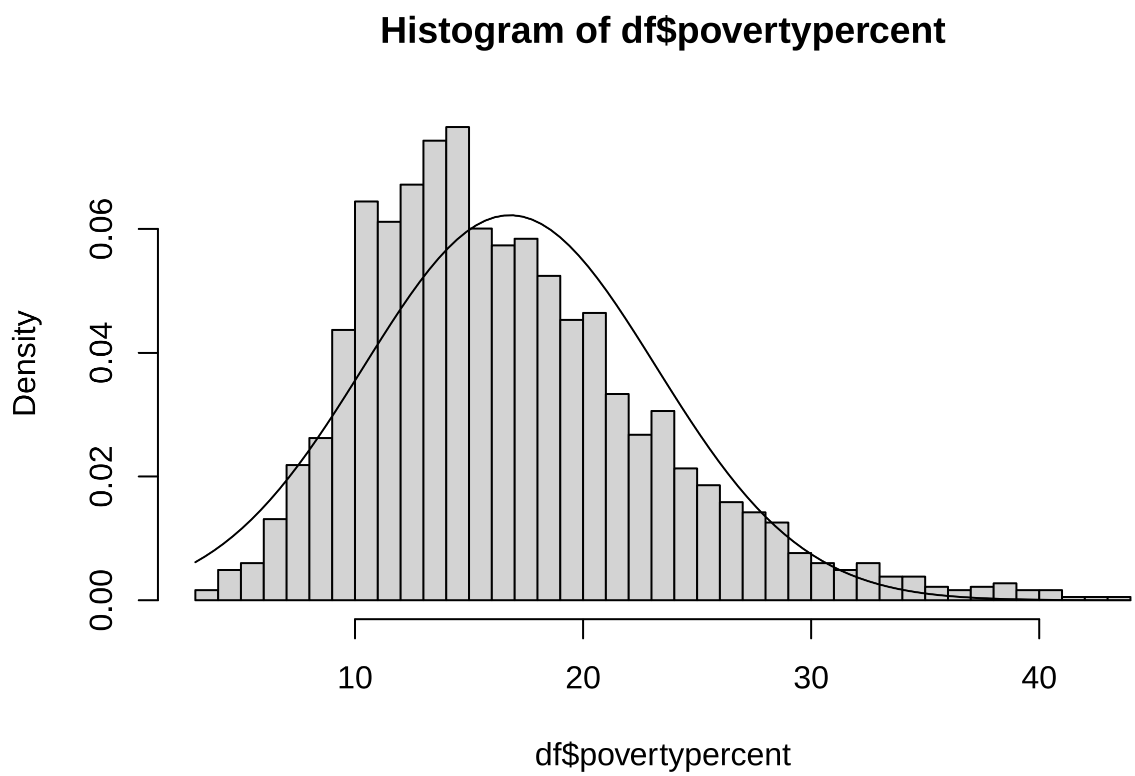
Variable 7 - povertypercent

Another continuous ratio variable not normally distributed with 0 missing values, 42 outliers (18 of them severe), all on the higher end. We create an additional ordinal factor “f.Pov%”.

```
summary(df$povertypercent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.70   12.15   15.70   16.79   20.40   44.00
```

```r
hist(df$povertypercent, breaks = 30, freq = F)
curve(dnorm(x, mean(df$povertypercent), sd(df$povertypercent)), add = T)
```



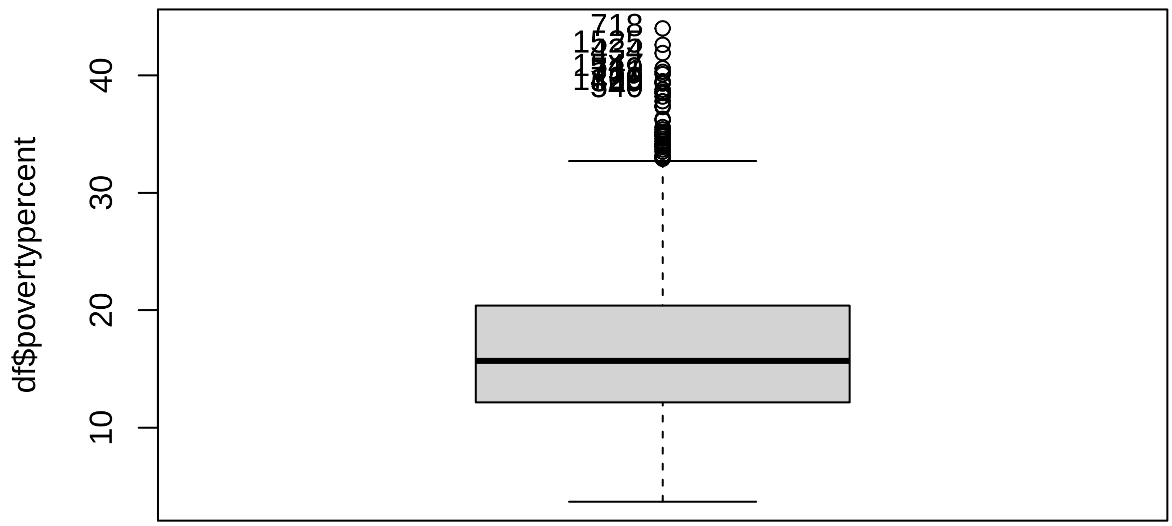

```r
shapiro.test(df$povertypercent)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$povertypercent
## W = 0.95557, p-value < 2.2e-16
```

```r
sum(is.na(df$povertypercent))
```

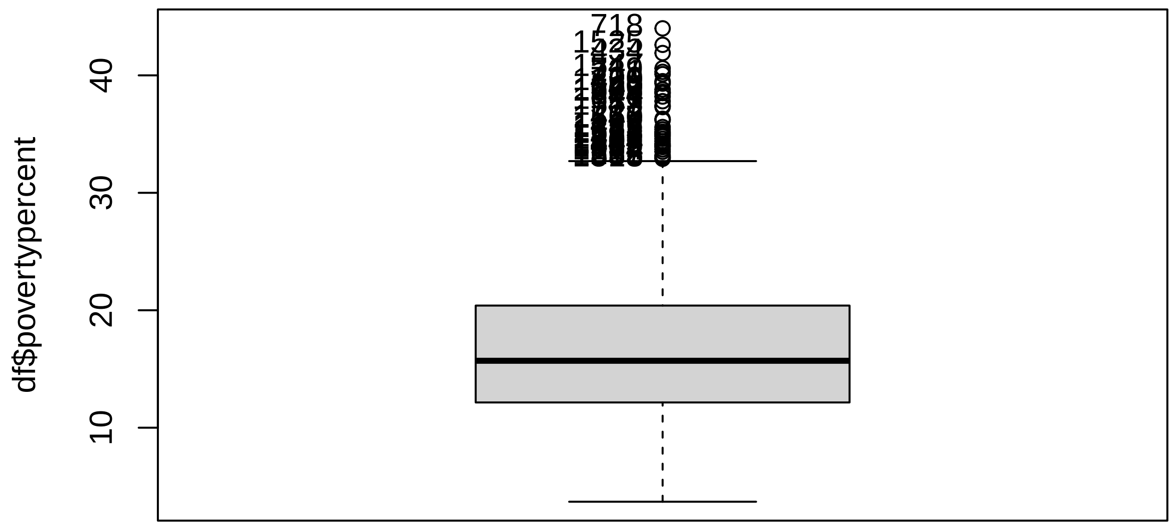
```
## [1] 0
```

```r
Boxplot(df$povertypercent)
```




```
## [1]  718 1525  434 1547  719  731  720 1468  329  540
```

```r
length(Boxplot(df$povertypercent, id = list(n=Inf)))
```




```
## [1] 42
```

```
sevout_povertypercent = (quantile(df$povertypercent,0.25)+(3*((quantile(df$povertypercent,0.75)-quantil
length(which(df$povertypercent > sevout_povertypercent))
```

```
## [1] 18
```

```
df$f.povertypercent <- ifelse(df$povertypercent <= 12.15, 1, ifelse(df$povertypercent > 12.15 & df$pove
df$f.povertypercent <- factor(df$f.povertypercent, labels=c("LowPov%","LowMidPov%","HighMidPov%","HighP
table(df$f.povertypercent)
```

```
##
##     LowPov%  LowMidPov% HighMidPov%    HighPov%
##         458         468         451         454
```

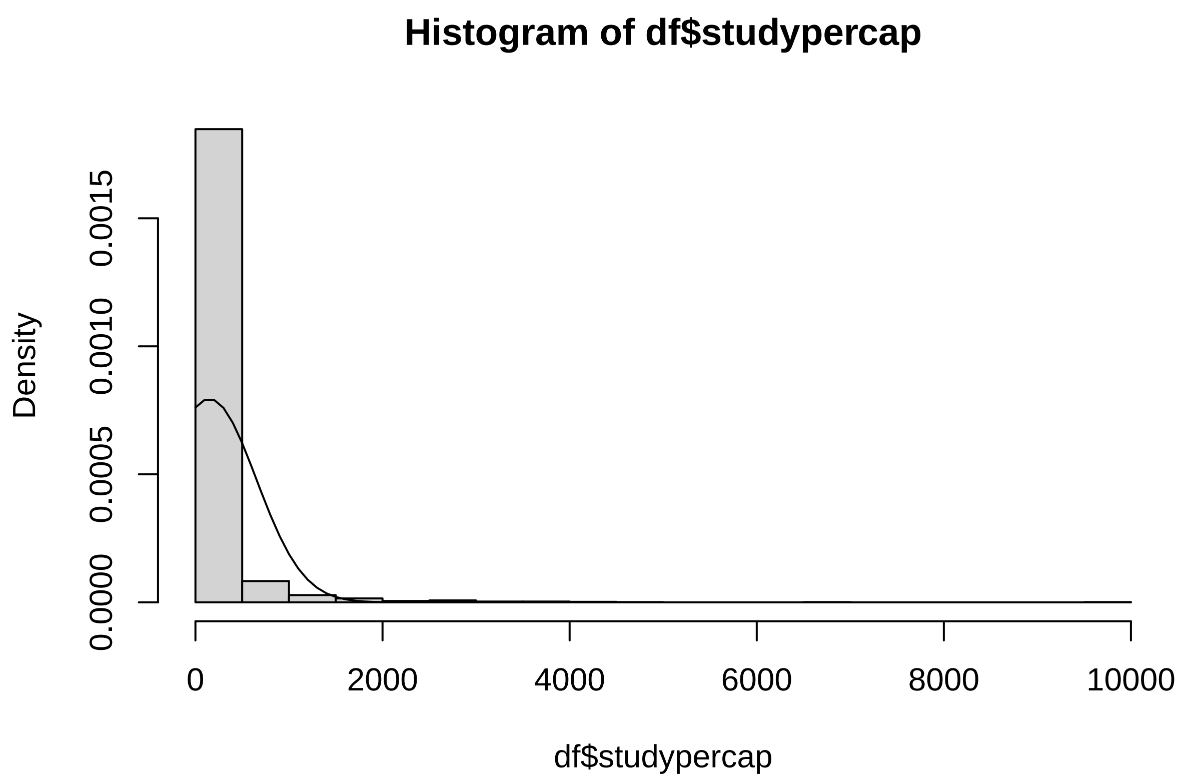
Variable 8 - studypercap

Another continuous ratio variable. This variable has the peculiarity of having a lot of 0s (median is also 0 so more than half of the counties don't perform cancer related clinical trials). It is not normally distributed and has 0 missing values, 307 outliers (281 of them severe), all on the higher end. We create an additional ordinal factor "f.studypercap" grouping the counties with 0 clinical trials and splitting the rest by half.

```r
summary(df$studypercap)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   148.2    76.0  9762.3
```

```r
hist(df$studypercap, breaks = 30, freq = F)
curve(dnorm(x, mean(df$studypercap), sd(df$studypercap)), add = T)
```



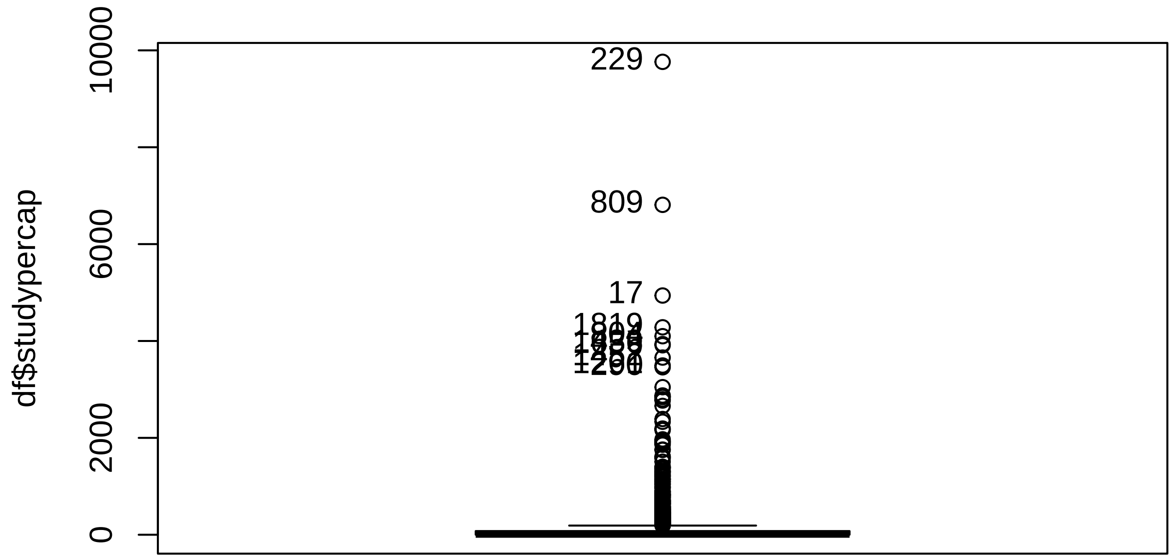

```r
shapiro.test(df$studypercap)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$studypercap
## W = 0.30754, p-value < 2.2e-16
```

```r
sum(is.na(df$studypercap))
```

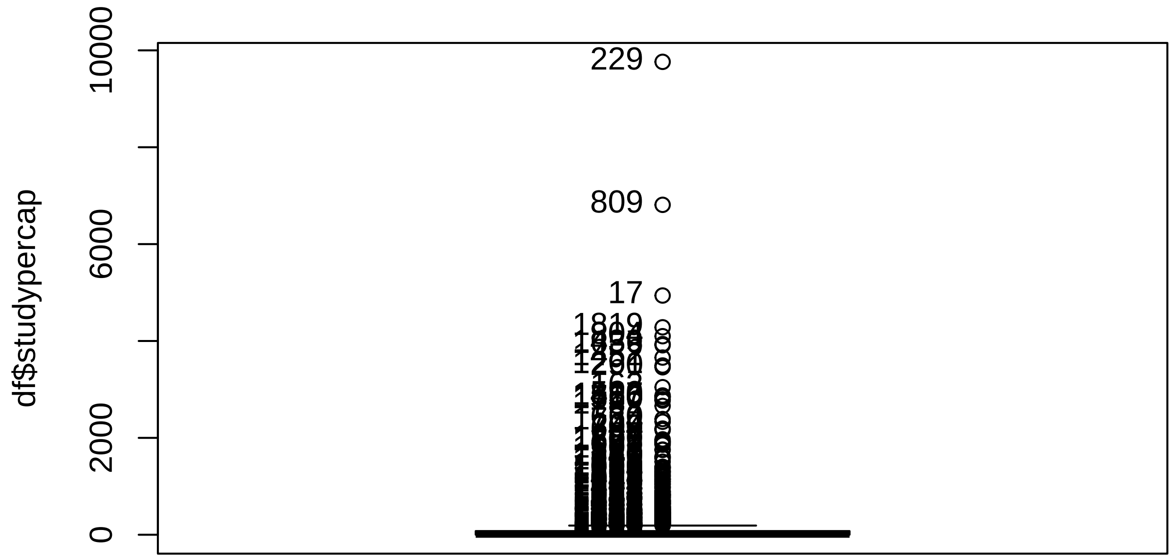
```
## [1] 0
```

```r
Boxplot(df$studypercap)
```




```
## [1]  229  809   17 1819  804 1439 1656 1452 1261  290
```

```r
length(Boxplot(df$studypercap, id = list(n=Inf)))
```




```
## [1] 307
```

```
sevout_studypercap = (quantile(df$studypercap,0.25)+(3*((quantile(df$studypercap,0.75)-quantile(df$study
length(which(df$studypercap > sevout_studypercap))
```

```
## [1] 281
```

```
studypercapNot0 <- df$studypercap[df$studypercap > 0]
summary(studypercapNot0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.03   57.18  162.13  405.62  422.18 9762.31
```

```
df$f.studypercap <- ifelse(df$studypercap == 0, 1, ifelse(df$studypercap > 0 & df$studypercap <= 162.13
df$f.studypercap <- factor(df$f.studypercap, labels=c("NoTrials","MidTrials","HighTrials"), order = T, l
table(df$f.studypercap)
```

```
##
##   NoTrials  MidTrials HighTrials
##       1162        334        335
```

Variable 9 - binnedinc

This is a string variable right now, but we can convert it to numerical by taking the midpoint in the bin as its value. Then we can treat it as a continuous ratio variable and analyze it. It has no missing values and the only outliers come from the same bin (the highest bin) which amount to 186 counties (all of them considered severe outliers). We create a factor variable "f.binnedinc" according to the quartiles.

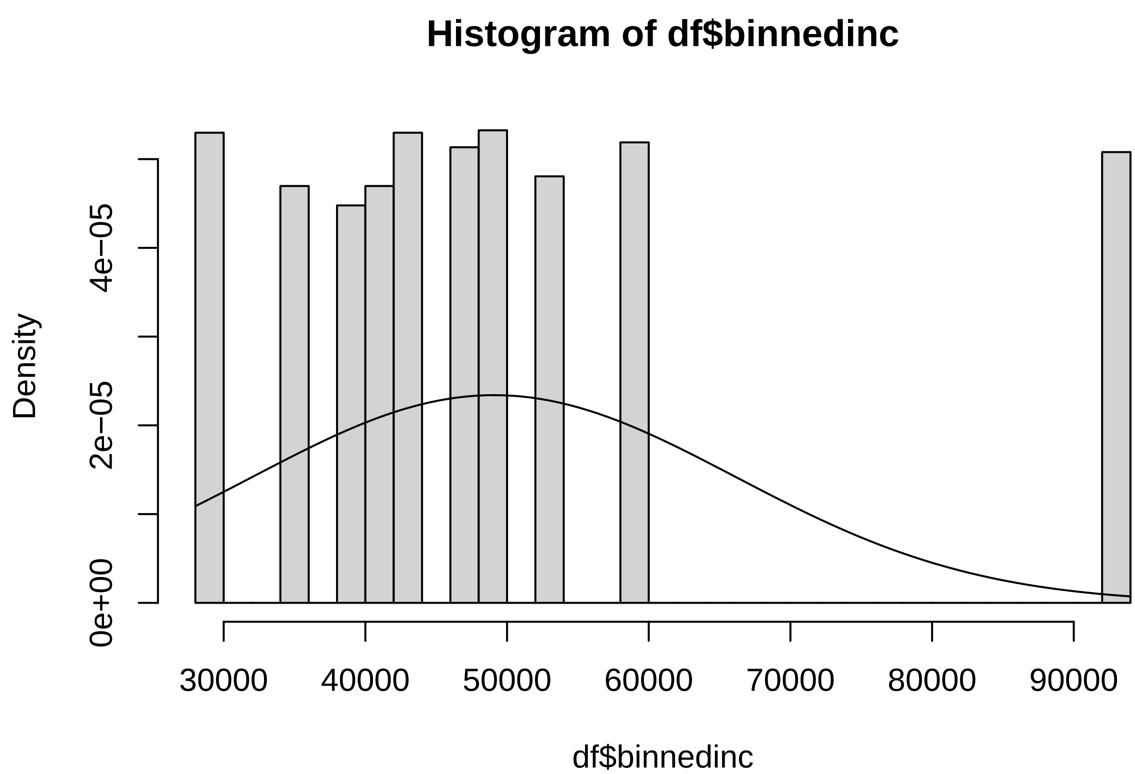
```
summary(df$binnedinc)
```

```
##    Length     Class      Mode
##      1831 character character
```

```
# Use regex to remove the [,],( and ) from the rows:
inc.midpoints.text <- gsub("[\\[\\]()]", "", df$binnedinc, perl = T)
# Separate them into two numbers
inc.midpoints.text.sep <- strsplit(inc.midpoints.text, ",")
# Convert them to numbers and apply a mean between them to find the midpoint
df$binnedinc <- sapply(inc.midpoints.text.sep, function(x) mean(as.numeric(x)))
summary(df$binnedinc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28429   38888   46611   49082   52796   93565
```

```
hist(df$binnedinc, breaks = 30, freq = F)
curve(dnorm(x, mean(df$binnedinc), sd(df$binnedinc)), add = T)
```



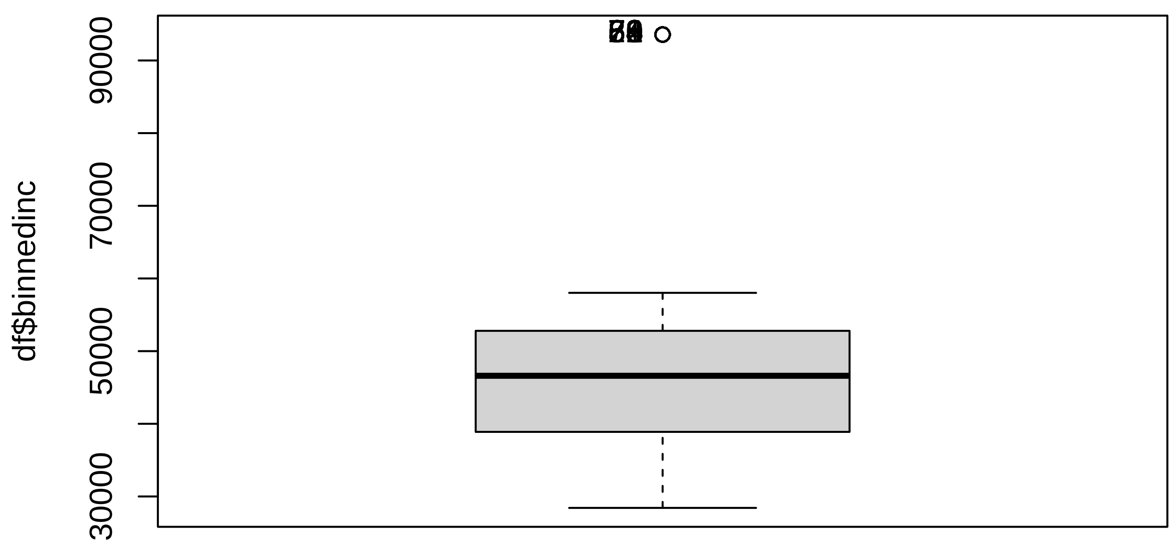
```r
shapiro.test(df$binnedinc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$binnedinc
## W = 0.79199, p-value < 2.2e-16
```

```r
sum(is.na(df$binnedinc))
```

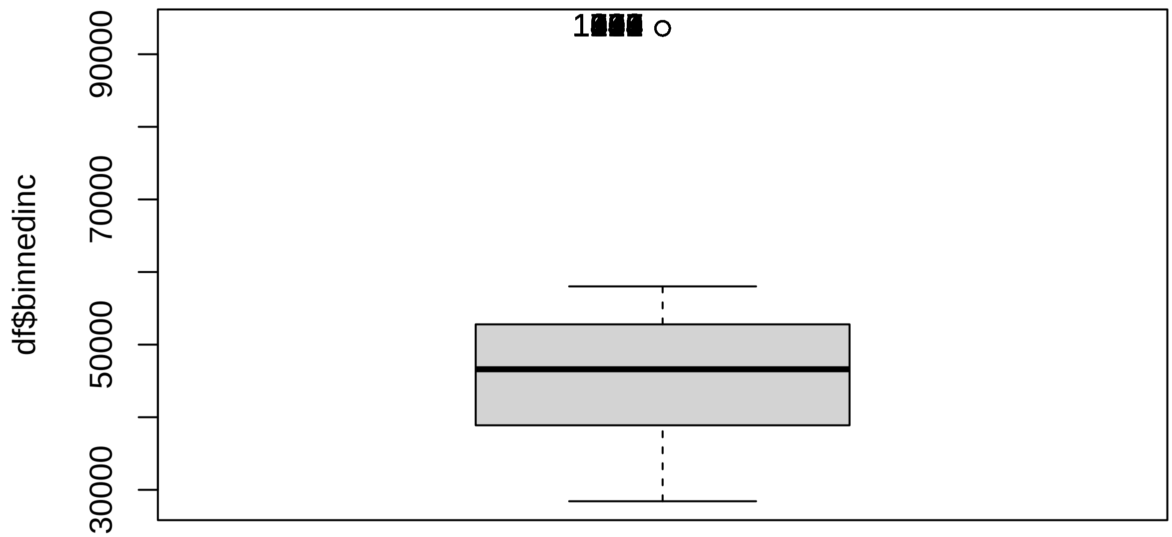
```
## [1] 0
```

```r
Boxplot(df$binnedinc)
```




```
##  [1]  8 26 50 54 63 69 71 72 73 83
```

```r
length(Boxplot(df$binnedinc, id = list(n=Inf)))
```




```
## [1] 186
```

```r
sevout_binnedinc = (quantile(df$binnedinc,0.25)+(3*((quantile(df$binnedinc,0.75)-quantile(df$binnedinc,0
length(which(df$binnedinc > sevout_binnedinc))
```

```
## [1] 186
```

```r
df$f.binnedinc <- ifelse(df$binnedinc <= 38888, 1, ifelse(df$binnedinc > 38888 & df$binnedinc <= 46611,
df$f.binnedinc <- factor(df$f.binnedinc, labels=c("LowIncPerCap","LowMidIncPerCap","HighMidIncPerCap","H
table(df$f.binnedinc)
```

```
##
##     LowIncPerCap  LowMidIncPerCap HighMidIncPerCap    HighIncPerCap
##              366              530              559              376
```

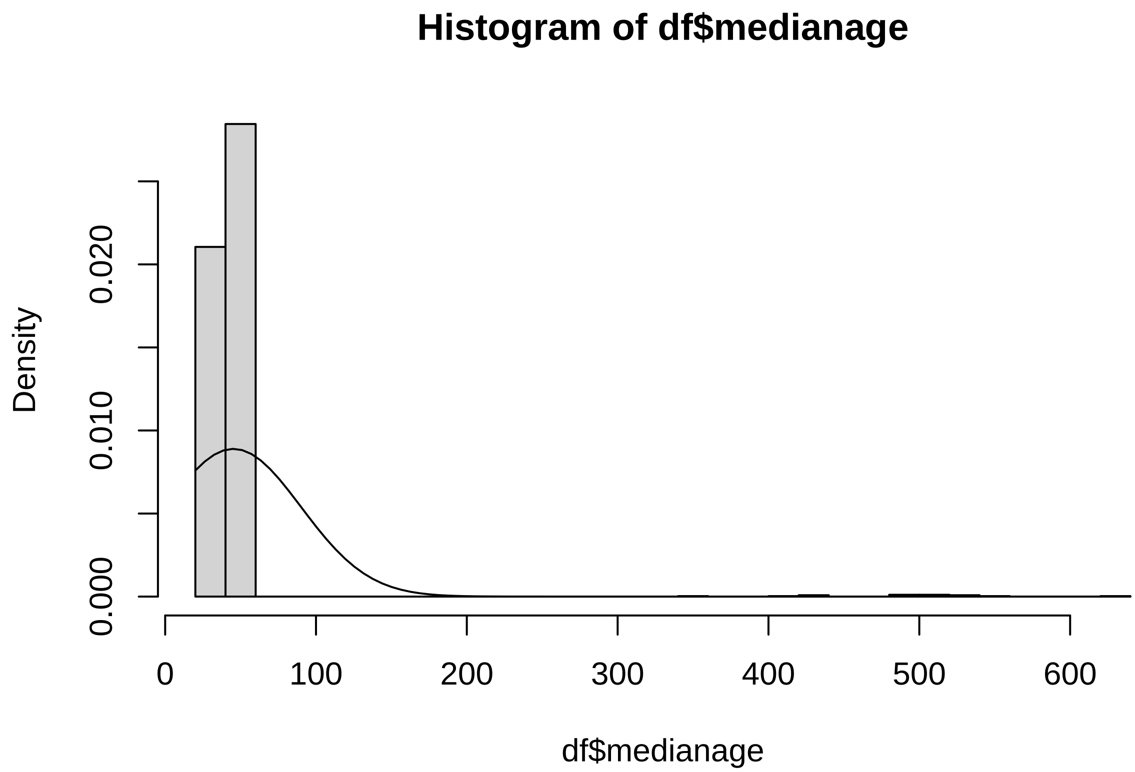
Variable 10 - medianage

This is a continuous interval variable. By using a histogram we see that there are some data points that make no sense (median ages over 100), so the data is erroneous. Since we have data for male median age and female median age will clean the data by replacing the ouliers by the mean of male and female age. After cleaning the data the variable has no missing data, is not normal by means of the shapiro test and has 50 outliers (5 of them severe) in both ends of the spectrum. We create a factor variable "f.medianage" according to the quartiles.

```
summary(df$medianage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.30   37.90   40.90   45.25   44.00  624.00
```

```
hist(df$medianage, breaks = 30, freq = F)
curve(dnorm(x, mean(df$medianage), sd(df$medianage)), add = T)
```

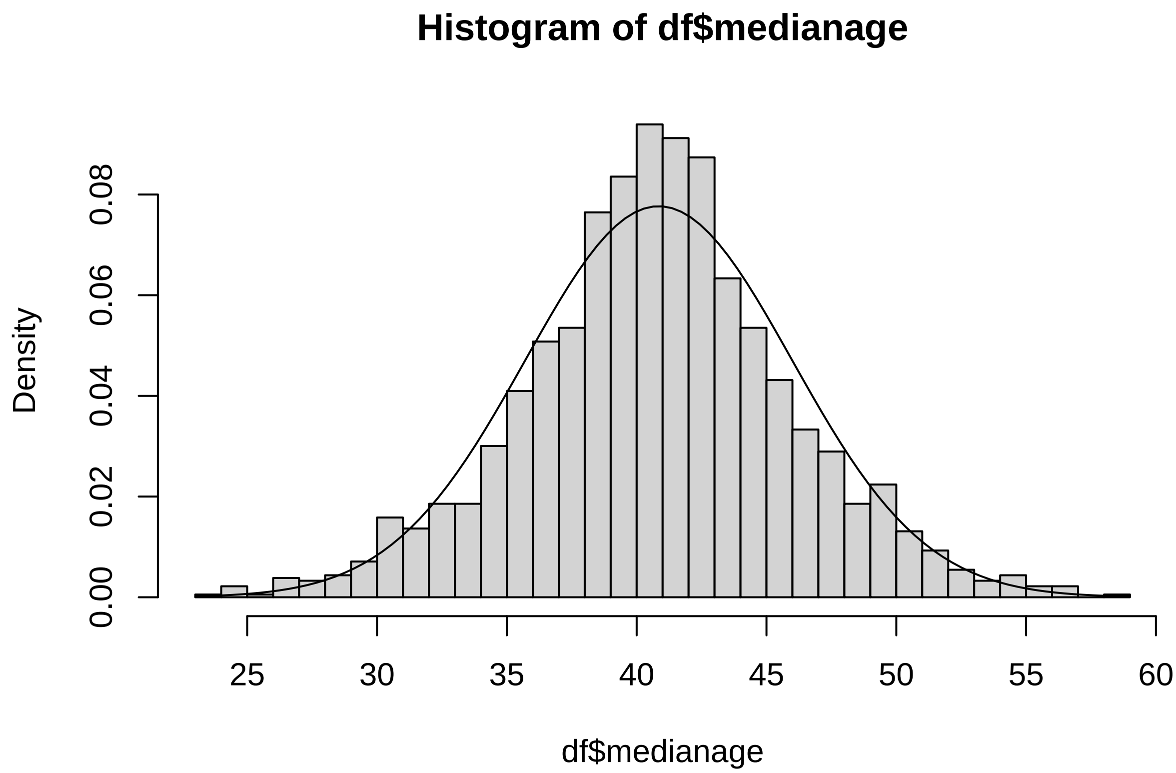



```
df$medianage[df$medianage>100] <- (df$medianagemale[df$medianage > 100] + df$medianagefemale[df$mediana

summary(df$medianage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.30   37.85   40.90   40.85   43.85   59.00
```

```
hist(df$medianage, breaks = 30, freq = F)
curve(dnorm(x, mean(df$medianage), sd(df$medianage)), add = T)
```



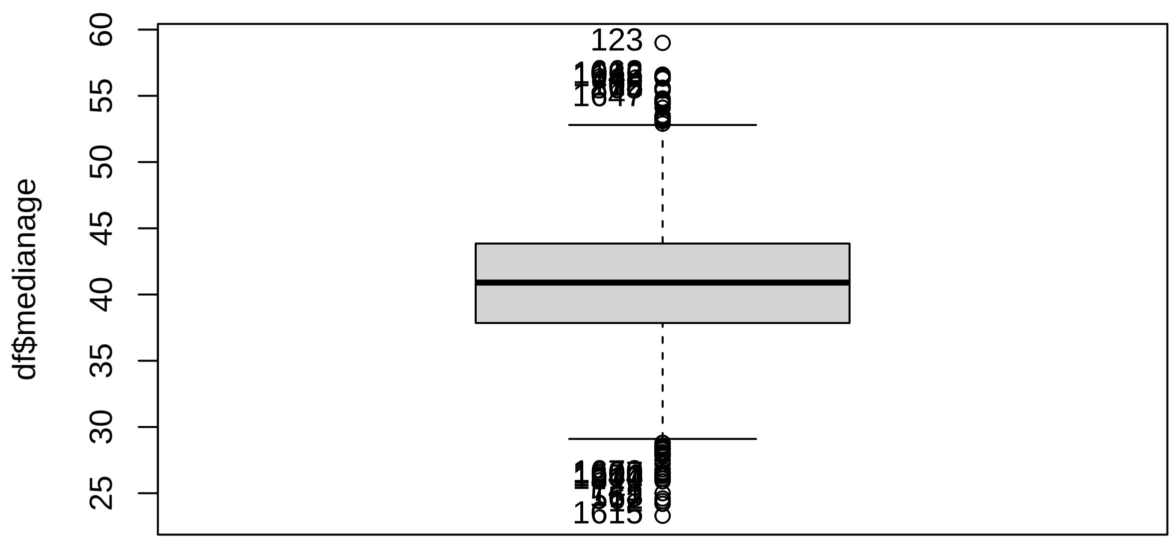

```r
shapiro.test(df$medianage)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$medianage
## W = 0.99506, p-value = 9.423e-06
```

```r
sum(is.na(df$medianage))
```

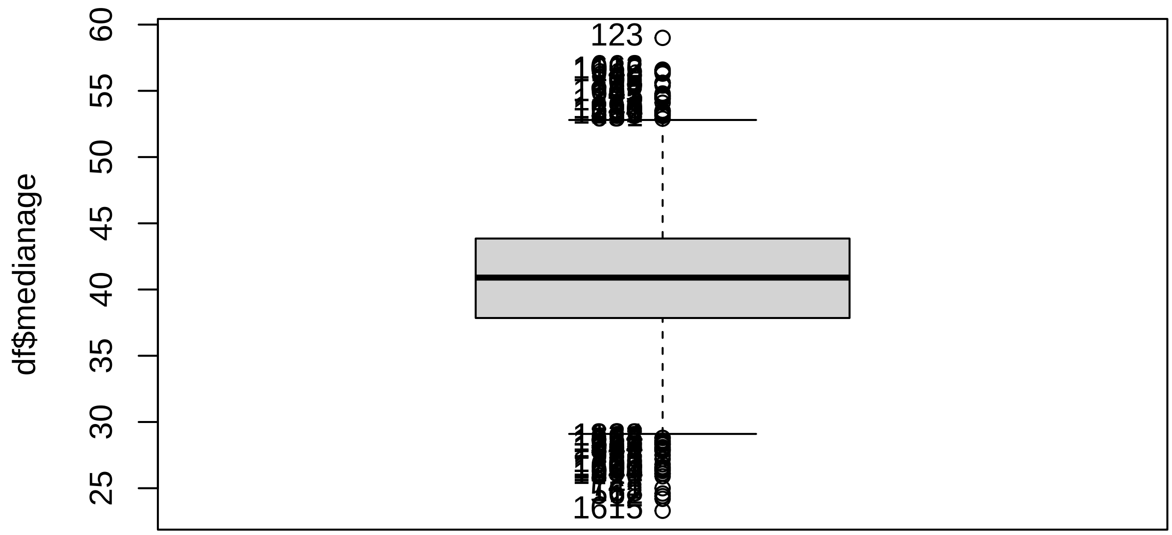
```
## [1] 0
```

```r
Boxplot(df$medianage)
```




```
##  [1] 1615   12  562  168  741 1810  254 1641 1607 1670  123  662 1016  632 1148
## [16]  112  178  208  865 1647
```

```r
length(Boxplot(df$medianage, id = list(n=Inf)))
```




```
## [1] 51
```

```
sevout_medianage = (quantile(df$medianage,0.25)+(3*((quantile(df$medianage,0.75)-quantile(df$medianage,0
length(which(df$medianage > sevout_medianage))
```

```
## [1] 5
```

```
df$f.medianage <- ifelse(df$medianage <= 37.85, 1, ifelse(df$medianage > 37.85 & df$medianage <= 40.90,
df$f.medianage <- factor(df$f.medianage, labels=c("LowAge","LowMidAge","HighMidAge","HighAge"), order =
table(df$f.medianage)
```

```
##
##     LowAge  LowMidAge HighMidAge    HighAge
##        458        466        460        447
```

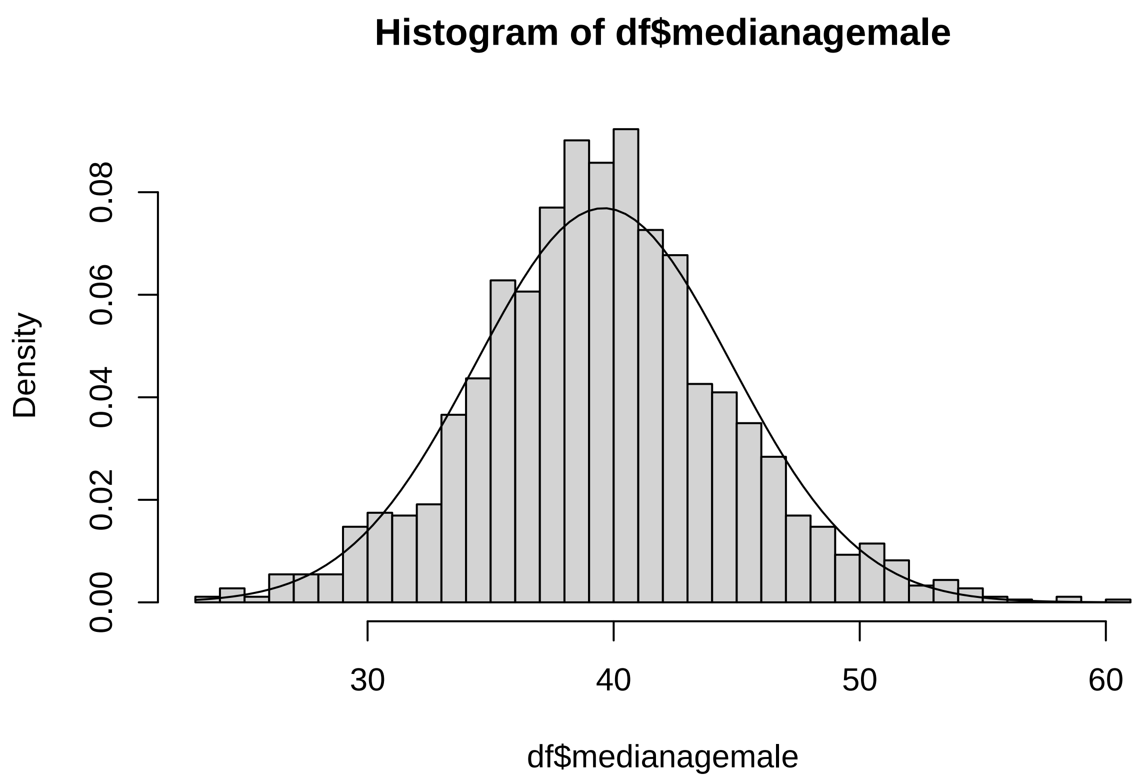
Variable 11 - medianagemale

Very similar to the previous variable, this is a continuous interval variable, but with no apparent erroneous input. The variable has no missing data, is not normal by means of the shapiro test and has 46 outliers (6 of them severe) in both ends of the spectrum. We create a factor variable "f.medianagemale" according to the quartiles. The summary shows that male median age is slightly lower than median age (and thus lower than female median age).

```r
summary(df$medianagemale)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.00   36.40   39.50   39.59   42.60   60.20
```

```r
hist(df$medianagemale, breaks = 30, freq = F)
curve(dnorm(x, mean(df$medianagemale), sd(df$medianagemale)), add = T)
```



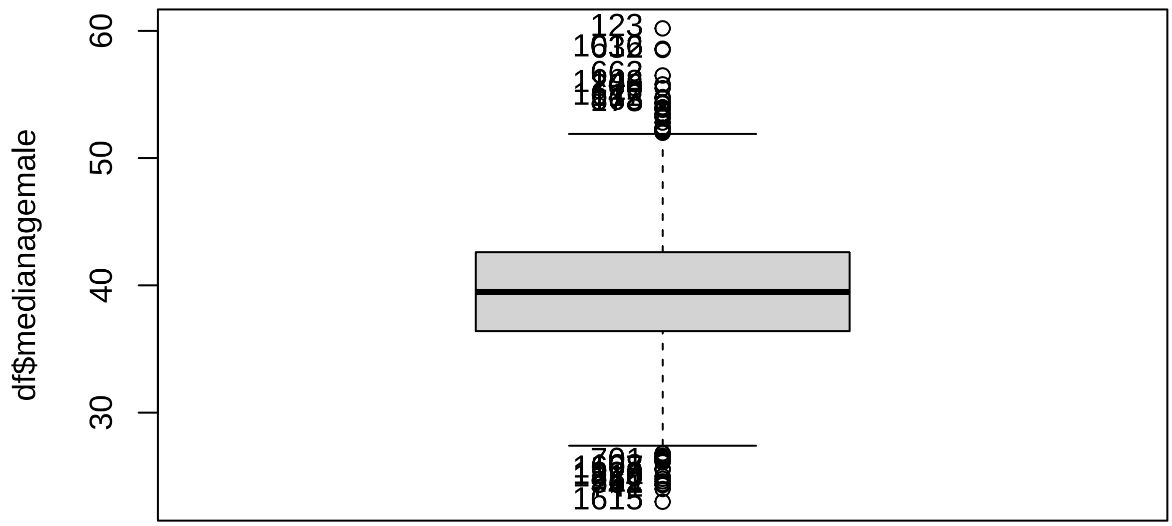

```r
shapiro.test(df$medianagemale)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$medianagemale
## W = 0.99404, p-value = 9.877e-07
```

```r
sum(is.na(df$medianagemale))
```

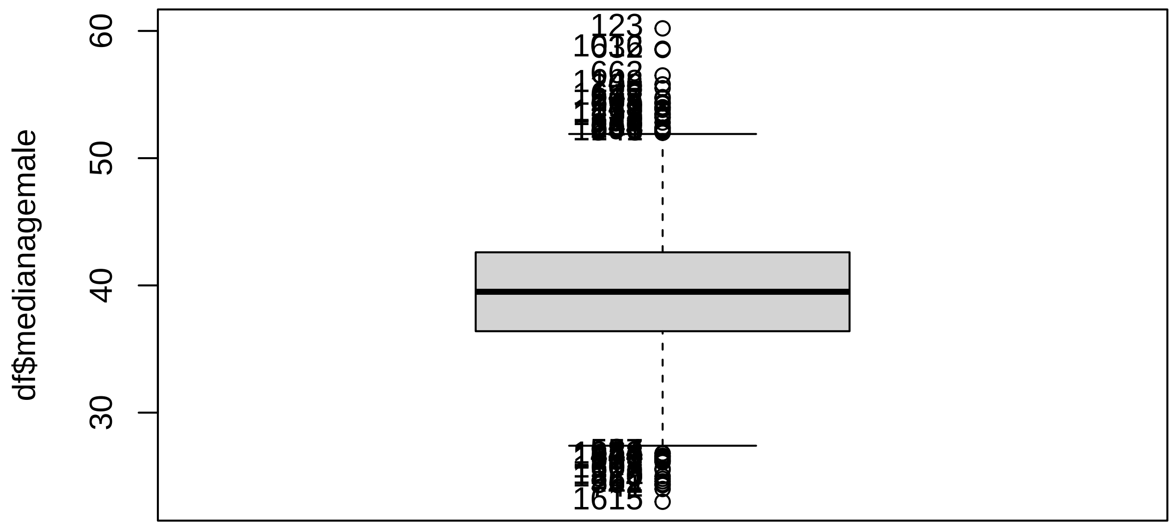
```
## [1] 0
```

```r
Boxplot(df$medianagemale)
```




```
##  [1] 1615  741  562   12  254 1810 1223 1607  168  701  123 1016  632  662 1148
## [16]  208 1647  112  865  178
```

```r
length(Boxplot(df$medianagemale, id = list(n=Inf)))
```




```
## [1] 46
```

```
sevout_medianagemale = (quantile(df$medianagemale,0.25)+(3*((quantile(df$medianagemale,0.75)-quantile(d
length(which(df$medianagemale > sevout_medianagemale))
```

```
## [1] 6
```

```
df$f.medianagemale <- ifelse(df$medianagemale <= 36.40, 1, ifelse(df$medianagemale > 36.40 & df$mediana
df$f.medianagemale <- factor(df$f.medianagemale, labels=c("LowAgeMale","LowMidAgeMale","HighMidAgeMale"
table(df$f.medianagemale)
```

```
##
##     LowAgeMale  LowMidAgeMale HighMidAgeMale    HighAgeMale
##            465            471            446            449
```

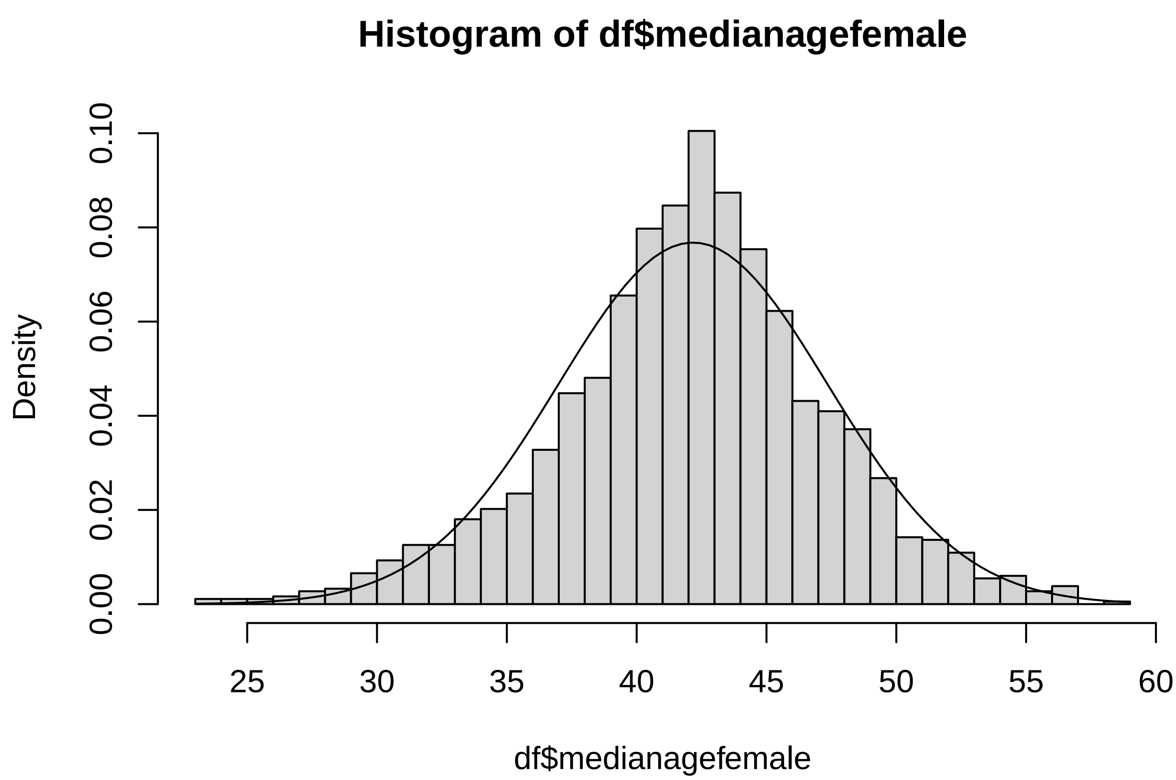
Variable 12 - medianagefemale

We repeat the analysis for female median age. The variable has no apparent erroneous input, no missing data, is not normal by means of the shapiro test and has 55 outliers (1 of them severe) in both ends of the spectrum. We create a factor variable "f.medianagefemale" according to the quartiles.

```
summary(df$medianagefemale)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.60   39.20   42.40   42.17   45.30   58.20
```

```r
hist(df$medianagefemale, breaks = 30, freq = F)
curve(dnorm(x, mean(df$medianagefemale), sd(df$medianagefemale)), add = T)
```



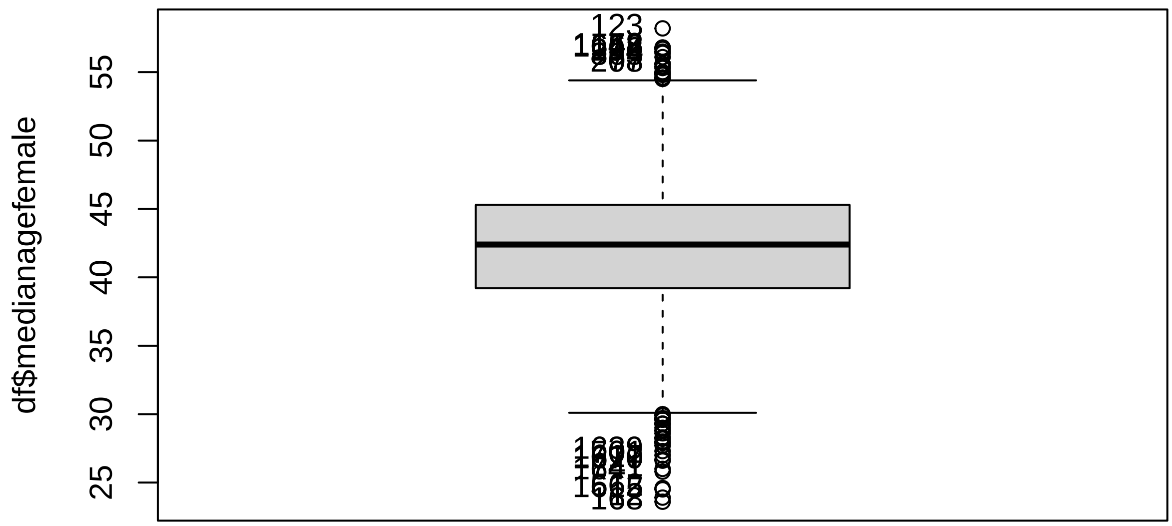

```r
shapiro.test(df$medianagefemale)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$medianagefemale
## W = 0.99321, p-value = 1.817e-07
```

```r
sum(is.na(df$medianagefemale))
```

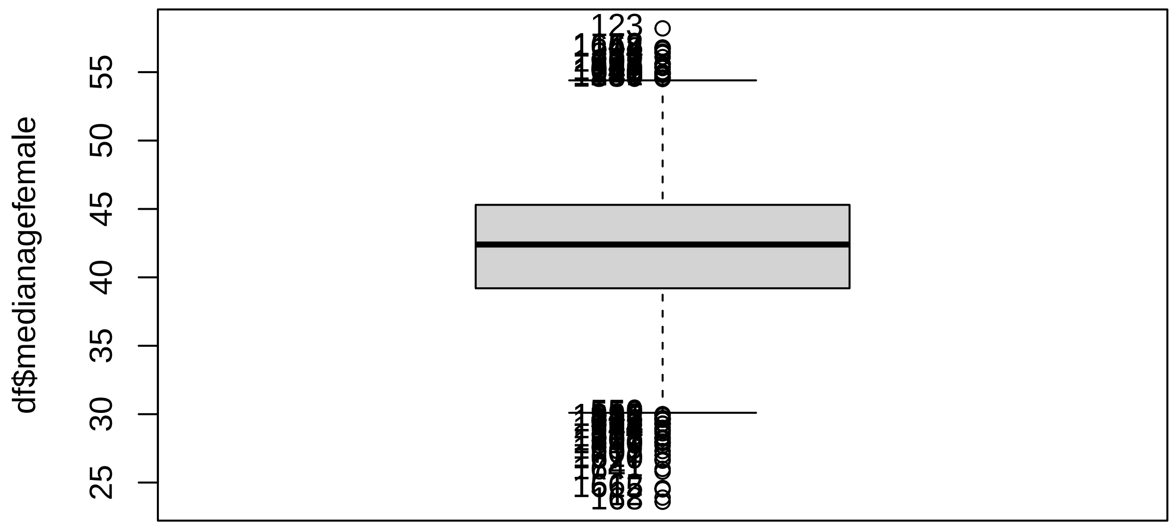
```
## [1] 0
```

```r
Boxplot(df$medianagefemale)
```




```
##  [1]  168   12 1615  562 1641  741 1670 1617  701 1639  123  178 1148  662 1658
## [16]  112  294  865   77  208
```

```r
length(Boxplot(df$medianagefemale, id = list(n=Inf)))
```



```
## [1] 55
```

```r
sevout_medianagefemale = (quantile(df$medianagefemale,0.25)+(3*((quantile(df$medianagefemale,0.75)-quan
length(which(df$medianagefemale > sevout_medianagefemale))
```

```
## [1] 1
```

```r
df$f.medianagefemale <- ifelse(df$medianagefemale <= 39.20, 1, ifelse(df$medianagefemale > 39.20 & df$me
df$f.medianagefemale <- factor(df$f.medianagefemale, labels=c("LowAgeFemale","LowMidAgeFemale","HighMid
table(df$f.medianagefemale)
```

```
##
##     LowAgeFemale  LowMidAgeFemale HighMidAgeFemale    HighAgeFemale
##              460              471              448              452
```

```r
summary(df$geography)
```

```
##    Length     Class      Mode
##      1831 character character
```

Variable 13 - geography

This is a string variable that is unique for each row of data. Since it is unique we could delete it, but it has info on not only the unique county of each observation, but also on its state. We will take this information and create a new variable named State that could be beneficial to our analysis. The new variable is a Nominal variable without missing values. However it has a lot of levels (50) with a few sparsly populated so it's not feasible to convert it to factor.

```r
sample(df$geography, 10)
```

```
##  [1] "Jackson County, Oregon"       "Cass County, North Dakota"
##  [3] "Montgomery County, Kansas"    "Fremont County, Wyoming"
##  [5] "Goshen County, Wyoming"       "Greene County, Virginia"
##  [7] "Roane County, West Virginia"  "Mifflin County, Pennsylvania"
##  [9] "Montcalm County, Michigan"    "Dubois County, Indiana"
```

```r
# Use regex to get the state (everything after the comma and white space):
df$state <- sub(".*,\\s*", "", df$geography)

summary(df$state)
```

```
##    Length     Class      Mode
##      1831 character character
```

```r
table(df$state)
```

```
##
##        Alabama         Alaska        Arizona       Arkansas     California
##             35             10              8             41             32
##       Colorado    Connecticut       Delaware        Florida        Georgia
##             34              7              1             38            100
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2             25             56             56             59
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##             61             75             40             10             14
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              8             51             51             59             66
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##             22             52             14              6             11
##     New Mexico       New York North Carolina   North Dakota           Ohio
##             20             41             62             32             49
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##             45             19             42              3             31
##   South Dakota      Tennessee          Texas           Utah        Vermont
##             39             60            136             18              7
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##             74             22             33             41             13
```

```r
unique(df$state)
```

```
##  [1] "Washington"     "West Virginia"  "Wisconsin"      "Nebraska"
##  [5] "Nevada"         "New Hampshire"  "New Jersey"     "New Mexico"
##  [9] "New York"       "Virginia"       "Michigan"       "Minnesota"
## [13] "North Carolina" "North Dakota"   "Alabama"        "Arkansas"
## [17] "California"     "Montana"        "Tennessee"      "Texas"
```

```
## [21] "Louisiana"      "Maine"          "Maryland"       "Massachusetts"
## [25] "Utah"           "Vermont"        "Colorado"       "Wyoming"
## [29] "Mississippi"    "Missouri"       "Kansas"         "Kentucky"
## [33] "Connecticut"    "Delaware"       "Florida"        "Oklahoma"
## [37] "Oregon"         "Ohio"           "Pennsylvania"   "Rhode Island"
## [41] "South Carolina" "Indiana"        "Iowa"           "Georgia"
## [45] "Hawaii"         "Idaho"          "Illinois"       "Alaska"
## [49] "Arizona"        "South Dakota"
```

```r
sum(is.na(df$state))
```

```
## [1] 0
```

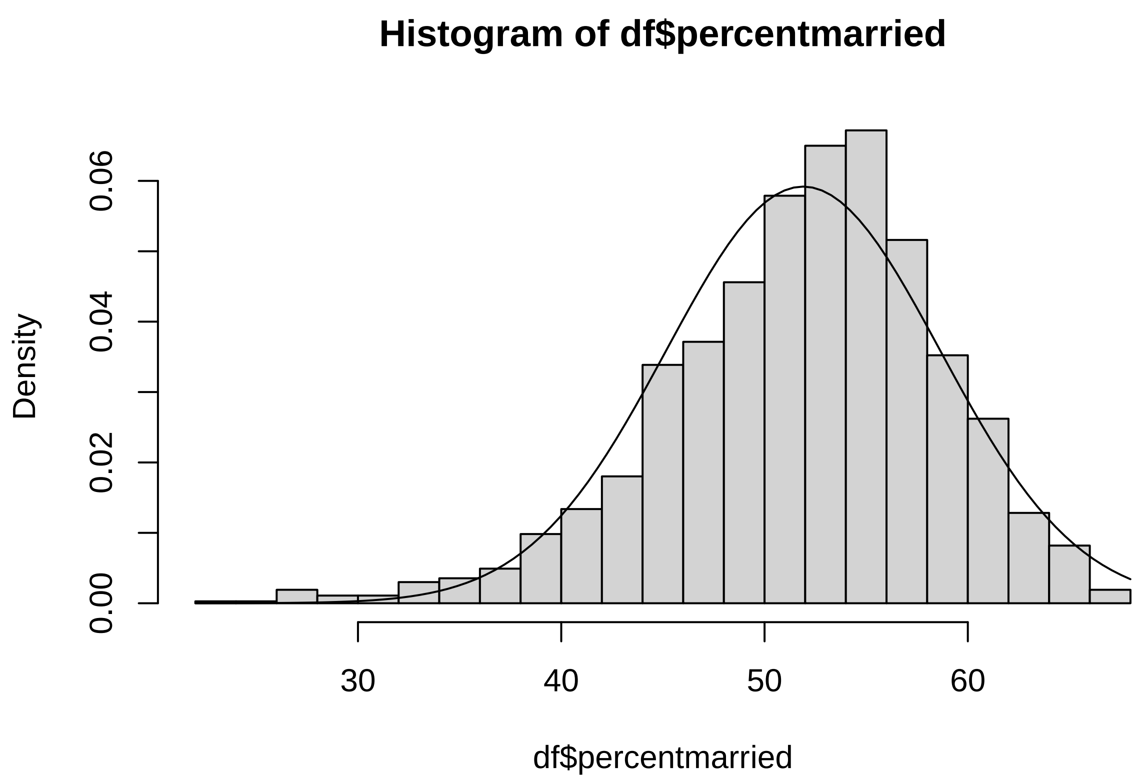
Variable 13 - percentmarried

Another continuous ratio variable not normally distributed with 0 missing values, 34 outliers (none of them severe), all on the lower end. We create an additional ordinal factor "f.percentmarried".

```r
summary(df$percentmarried)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    23.1    47.8    52.5    51.9    56.4    68.0
```

```r
hist(df$percentmarried, breaks = 30, freq = F)
curve(dnorm(x, mean(df$percentmarried), sd(df$percentmarried)), add = T)
```



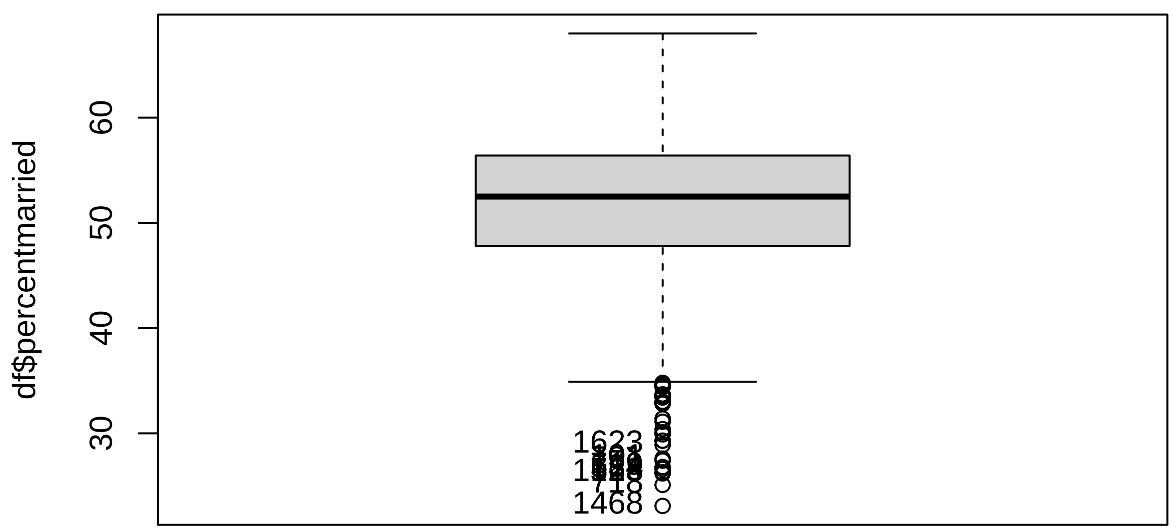
```r
shapiro.test(df$percentmarried)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$percentmarried
## W = 0.97753, p-value = 2.346e-16
```

```r
sum(is.na(df$percentmarried))
```

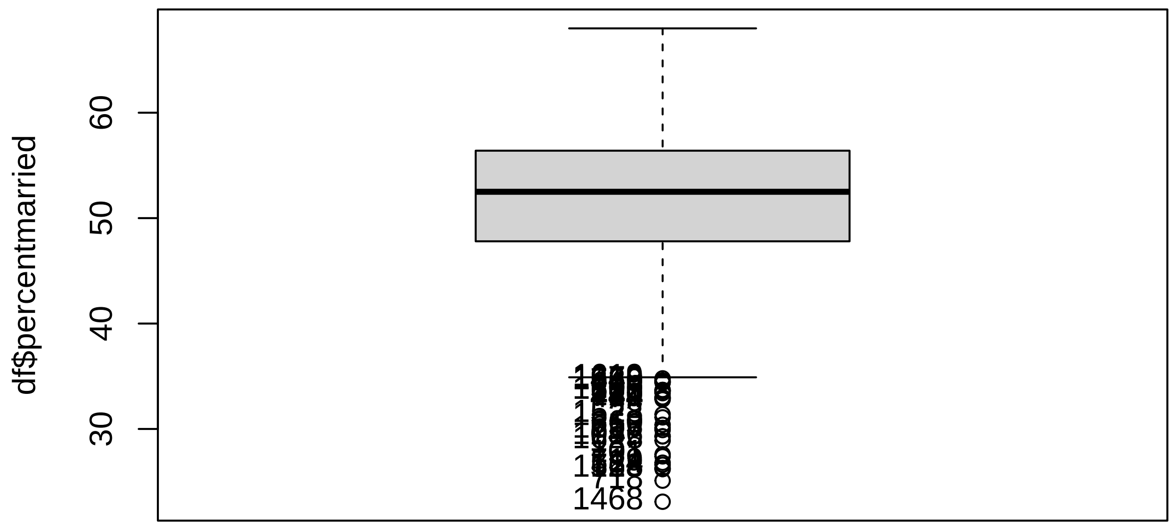
```
## [1] 0
```

```r
Boxplot(df$percentmarried)
```




```
##  [1] 1468  718  168 1525  723  534  719  731  101 1623
```

```r
length(Boxplot(df$percentmarried, id = list(n=Inf)))
```




```
## [1] 34
```

```
sevout_percentmarried = (quantile(df$percentmarried,0.25)+(3*((quantile(df$percentmarried,0.75)-quantil
length(which(df$percentmarried > sevout_percentmarried))
```

```
## [1] 0
```

```
df$f.percentmarried <- ifelse(df$percentmarried <= 47.8, 1, ifelse(df$percentmarried > 47.8 & df$percen
df$f.percentmarried <- factor(df$f.percentmarried, labels=c("LowMarriage%","LowMidMarriage%","HighMidMa
table(df$f.percentmarried)
```

```
##
##     LowMarriage%  LowMidMarriage% HighMidMarriage%    HighMarriage%
##              460              459              455              457
```

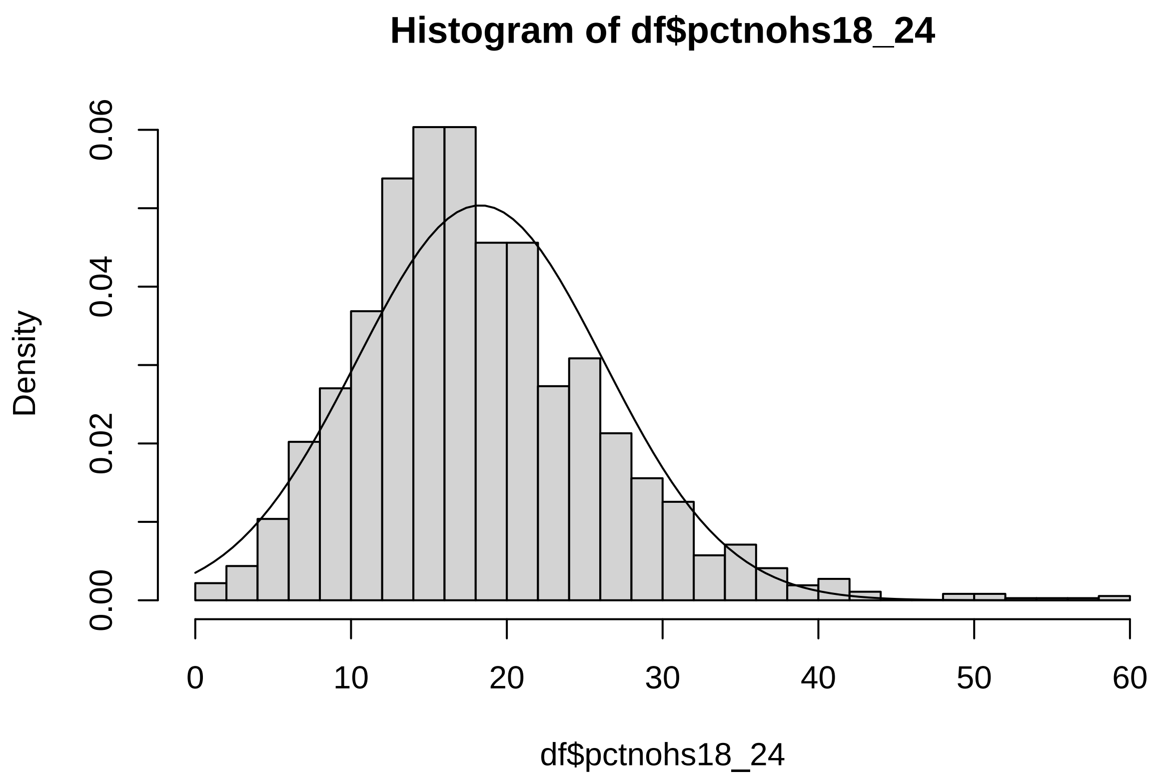
Variable 14 - pctnohs18_24

Another continuous ratio variable not normally distributed with 0 missing values, 34 outliers (none of them severe), all on the higher end. We create an additional ordinal factor "f.pctnohs18_24".

```
summary(df$pctnohs18_24)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.50   12.90   17.20   18.29   22.70   59.10
```

```r
hist(df$pctnohs18_24, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctnohs18_24), sd(df$pctnohs18_24)), add = T)
```



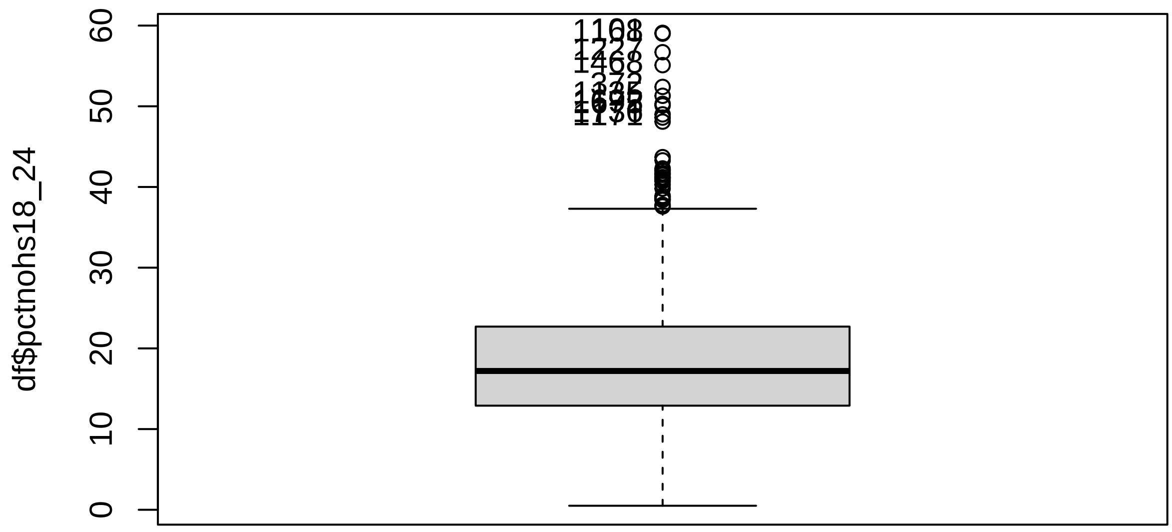

```r
shapiro.test(df$pctnohs18_24)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctnohs18_24
## W = 0.96205, p-value < 2.2e-16
```

```r
sum(is.na(df$pctnohs18_24))
```

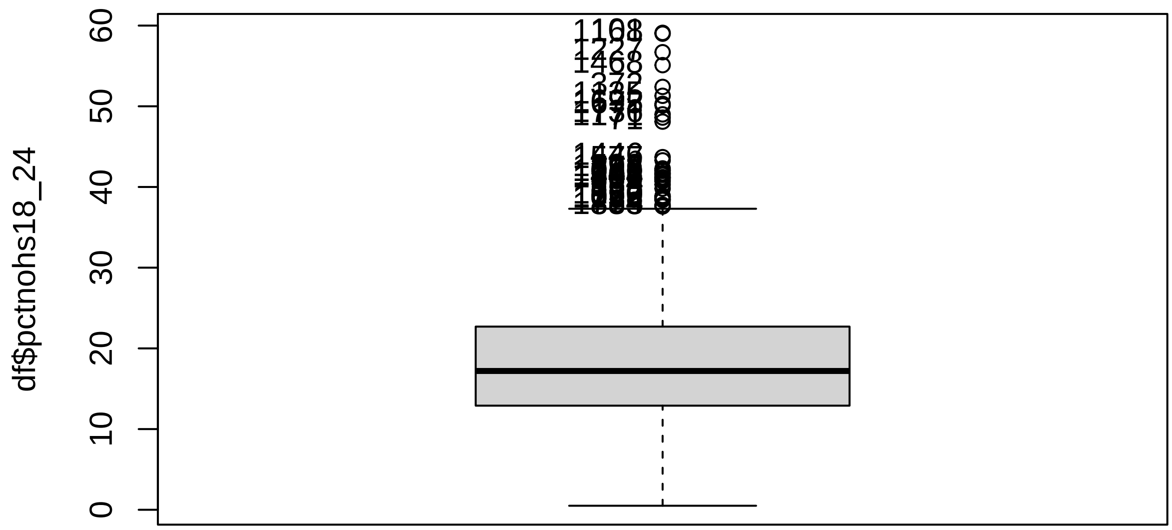
```
## [1] 0
```

```r
Boxplot(df$pctnohs18_24)
```




```
## [1]  101 1168 1227 1468  372 1135 1692 1675 1736 1171
```

```r
length(Boxplot(df$pctnohs18_24, id = list(n=Inf)))
```




```
## [1] 35
```

```r
sevout_pctnohs18_24 = (quantile(df$pctnohs18_24,0.25)+(3*((quantile(df$pctnohs18_24,0.75)-quantile(df$pc
length(which(df$pctnohs18_24 > sevout_pctnohs18_24))
```

```
## [1] 13
```

```r
df$f.pctnohs18_24 <- ifelse(df$pctnohs18_24 <= 12.90, 1, ifelse(df$pctnohs18_24 > 12.90 & df$pctnohs18_2
df$f.pctnohs18_24 <- factor(df$f.pctnohs18_24, labels=c("LowNoHighsc%","LowMidNoHighsc%","HighMidNoHighs
table(df$f.pctnohs18_24)
```

```
##
##     LowNoHighsc%  LowMidNoHighsc% HighMidNoHighsc%    HighNoHighsc%
##              459              461              455              456
```

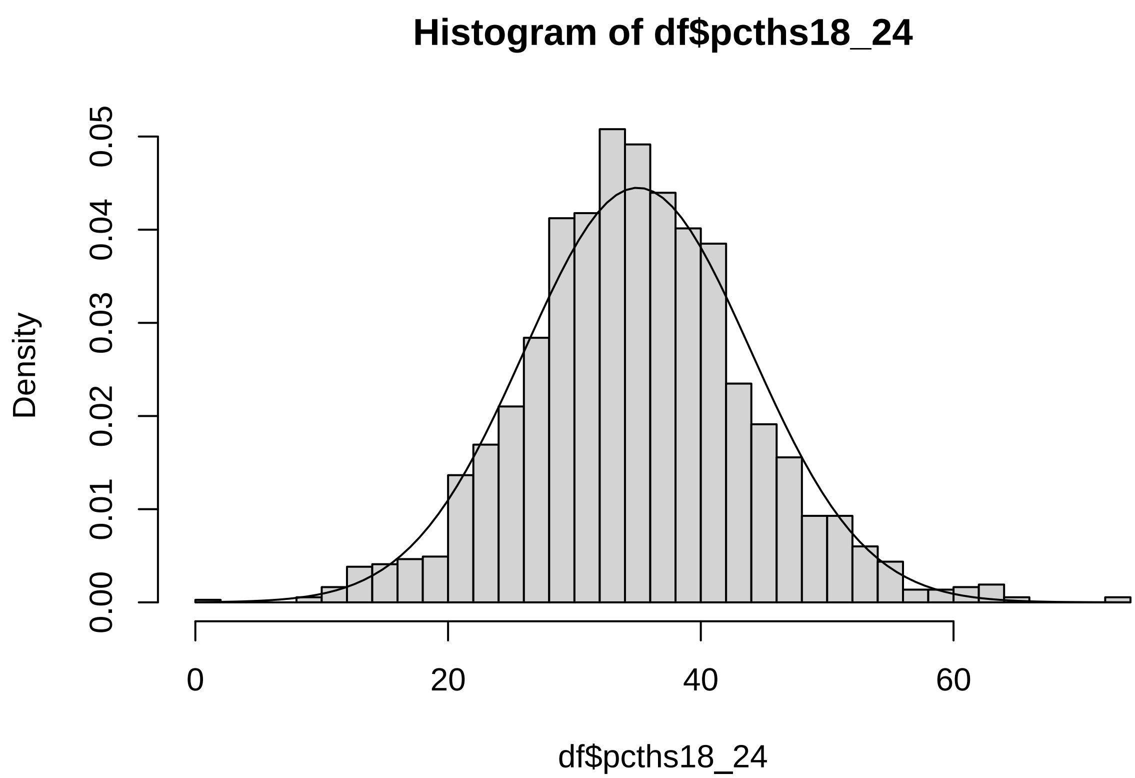
Variable 15 - pcths18_24

Another continuous ratio variable (related to the previous one) not normally distributed with 0 missing values, 33 outliers (9 of them severe) on both ends. There is one really severe outlier with 0 percent of High School Graduates, Greeley County, Kansas. It also has only 4.8% non High School Graduates (really low) and NA college graduates with a population of 1330. It seems like the values are probably false. For now we will leave it as such and later we will see how to deal with it. We create an additional ordinal factor "f.pcths18_24".

```r
summary(df$pcths18_24)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    29.2    34.7    35.0    40.5    72.5
```

```r
hist(df$pcths18_24, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pcths18_24), sd(df$pcths18_24)), add = T)
```



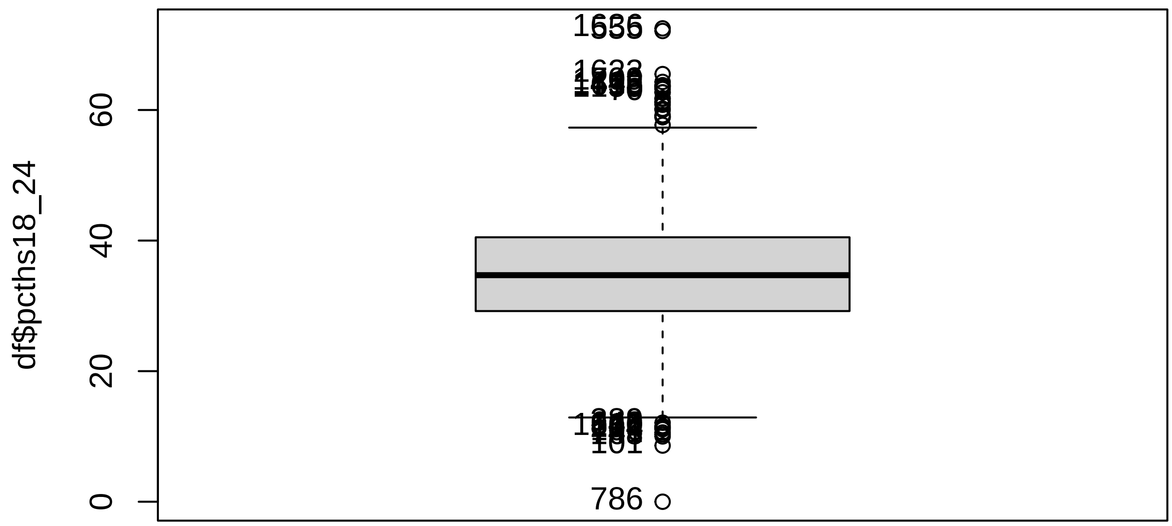

```r
shapiro.test(df$pcths18_24)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pcths18_24
## W = 0.99323, p-value = 1.922e-07
```

```r
sum(is.na(df$pcths18_24))
```

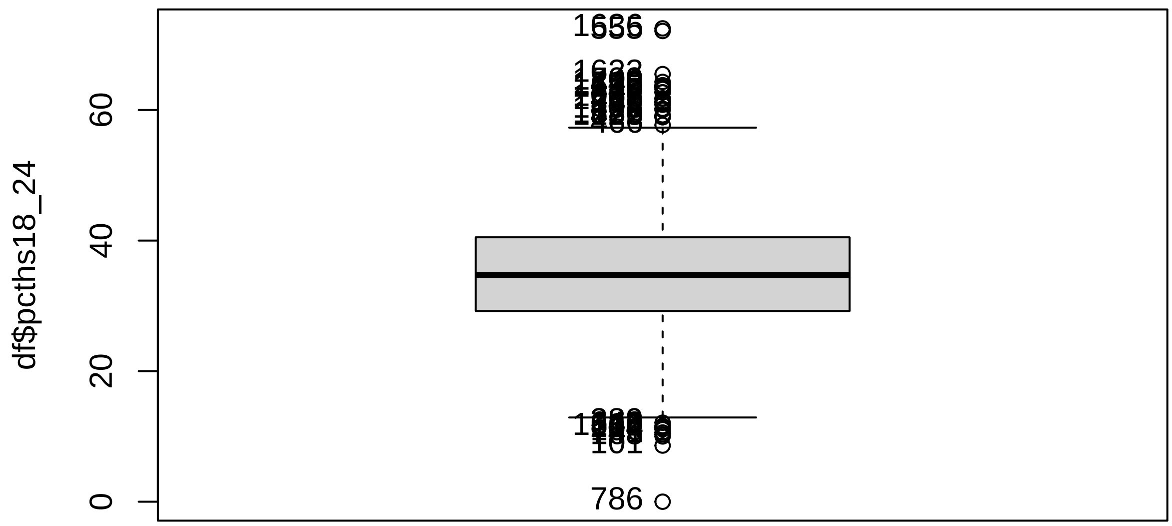
```
## [1] 0
```

```r
Boxplot(df$pcths18_24)
```




```
## [1]  101  106  131  168  294  389  642  741  786 1810 1636  555 1623 1709  443
## [16] 1716  436 1699 1155   70
```

```
length(Boxplot(df$pcths18_24, id = list(n=Inf)))
```




```
## [1] 33
```

```
sevout_pcths18_24 = (quantile(df$pcths18_24,0.25)+(3*((quantile(df$pcths18_24,0.75)-quantile(df$pcths18_
length(which(df$pcths18_24 > sevout_pcths18_24))
```

```
## [1] 9
```

```
df[786,]
```

```
##     avganncount avgdeathsperyear target_deathrate incidencerate medincome
## 786    1962.668                3            156.9      453.5494     52795
##     popest2015 povertypercent studypercap binnedinc medianage medianagemale
## 786       1330           10.8           0     52796      49.4          48.7
##     medianagefemale              geography percentmarried pctnohs18_24
## 786            49.9 Greeley County, Kansas           66.6          4.8
##     pcths18_24 pctsomecol18_24 pctbachdeg18_24 pcths25_over pctbachdeg25_over
## 786          0              NA            40.3         30.3              20.4
##     pctemployed16_over pctunemployed16_over pctprivatecoverage
## 786               60.5                  2.1               81.9
##     pctprivatecoveragealone pctempprivcoverage pctpubliccoverage
## 786                    60.5               42.7              28.8
##     pctpubliccoveragealone pctwhite  pctblack  pctasian pctotherrace
## 786                   10.5  87.1732 0.8986928 0.3267974     8.905229
##     pctmarriedhouseholds birthrate f.avganncount f.avgdeathsperyear
```

```
## 786             64.68172  5.687204 HighCaseCount       LowMortCount
##      f.deathrate    f.incidencerate   f.medincome f.popest2015 f.povertypercent
## 786 LowDeathrate HighMidDiagnPerCap HighMedianInc       LowPop          LowPov%
##     f.studypercap      f.binnedinc f.medianage f.medianagemale
## 786      NoTrials HighMidIncPerCap     HighAge     HighAgeMale
##     f.medianagefemale  state f.percentmarried f.pctnohs18_24
## 786     HighAgeFemale Kansas    HighMarriage%   LowNoHighsc%
```

```
df$f.pcths18_24 <- ifelse(df$pcths18_24 <= 29.2, 1, ifelse(df$pcths18_24 > 29.2 & df$pcths18_24 <= 34.7
df$f.pcths18_24 <- factor(df$f.pcths18_24, labels=c("LowHighsc%","LowMidHighsc%","HighMidHighsc%","High
table(df$f.pcths18_24)
```

```
##
##     LowHighsc%  LowMidHighsc% HighMidHighsc%    HighHighsc%
##            461            463            456            451
```

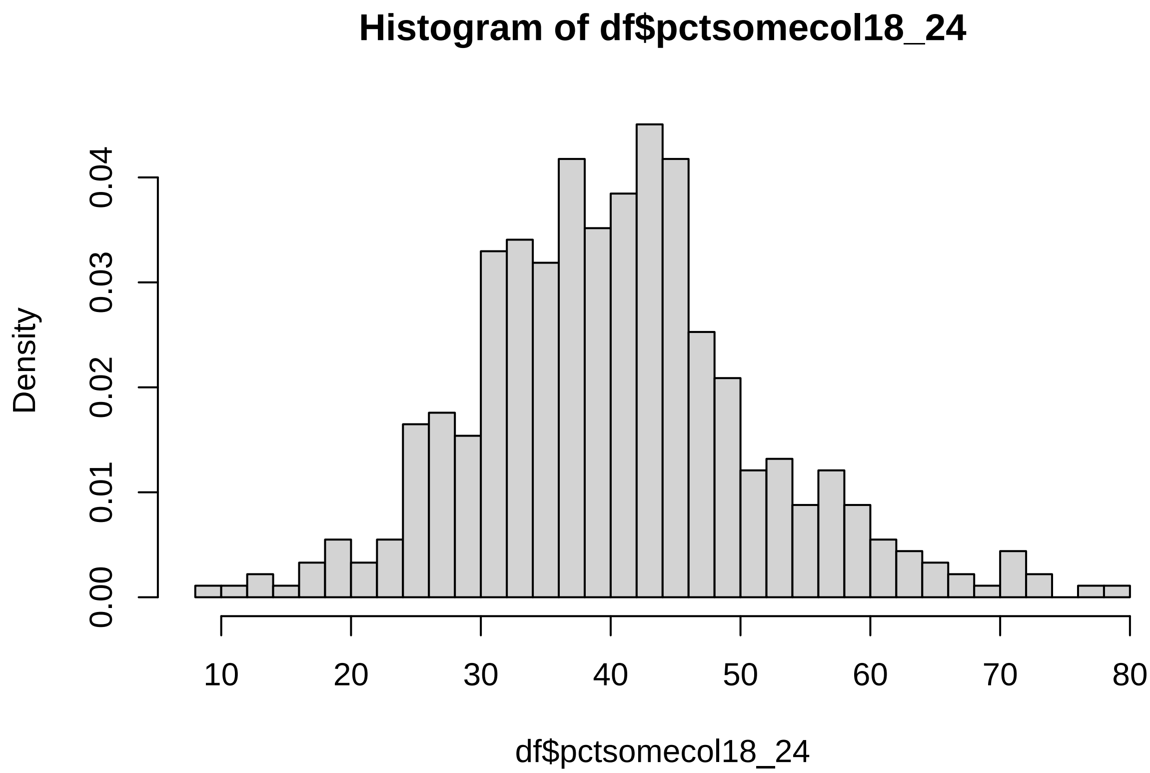
Variable 16 - pctsomecol18_24

Another continuous ratio variable (related to the 2 previous ones). It has 1376 missing values which is more than 75% of our sample. This is too much and we will take the decision to take this variable out of the study because of with such a high proportion of missing data, it will not provide meaningful information.

```
summary(df$pctsomecol18_24)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    9.60   33.25   40.10   40.48   46.10   78.30    1376
```

```
hist(df$pctsomecol18_24, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctsomecol18_24), sd(df$pctsomecol18_24)), add = T)
```




```r
sum(is.na(df$pctsomecol18_24))
```

```
## [1] 1376
```

```r
1376/1831*100
```

```
## [1] 75.15019
```

```r
#Removing the column
df <- subset(df, select = -pctsomecol18_24)
```

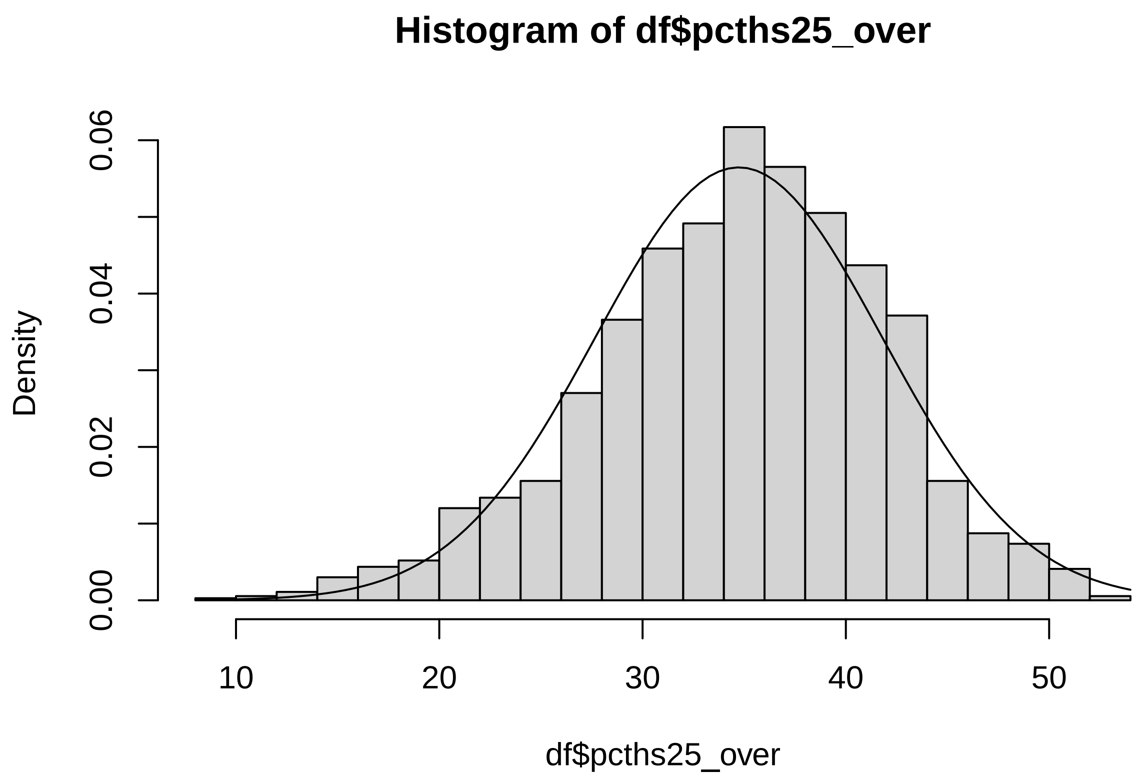
Variable 17 - pcths25_over

Another continuous ratio variable not normally distributed with 0 missing values, 18 outliers (none of them severe), all on the lower end. We create an additional ordinal factor "f.pcths25_over".

```r
summary(df$pcths25_over)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.30   30.35   35.30   34.73   39.65   52.70
```

```r
hist(df$pcths25_over, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pcths25_over), sd(df$pcths25_over)), add = T)
```



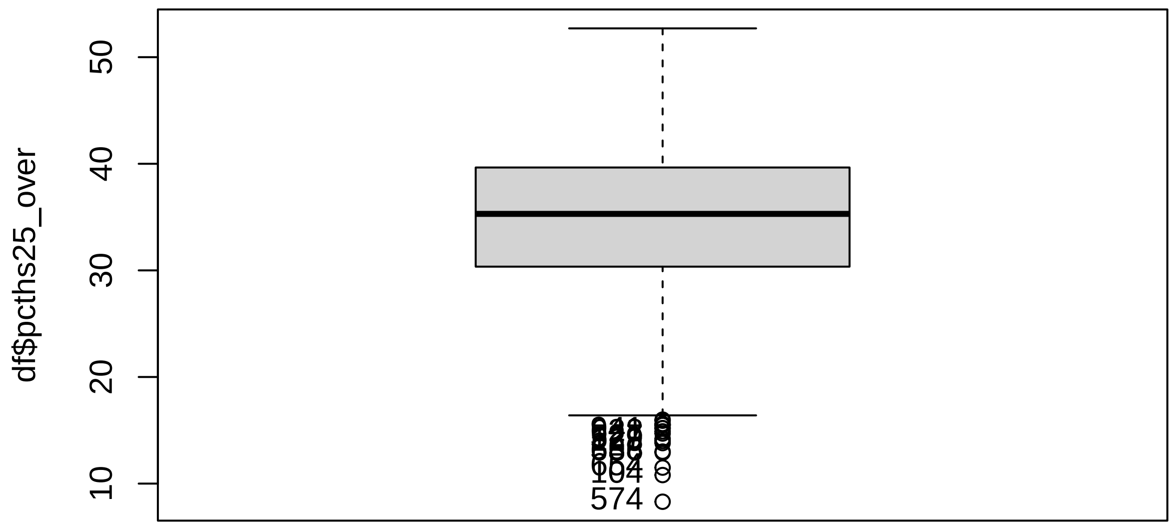

```r
shapiro.test(df$pcths25_over)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pcths25_over
## W = 0.99107, p-value = 3.741e-09
```

```r
sum(is.na(df$pcths25_over))
```

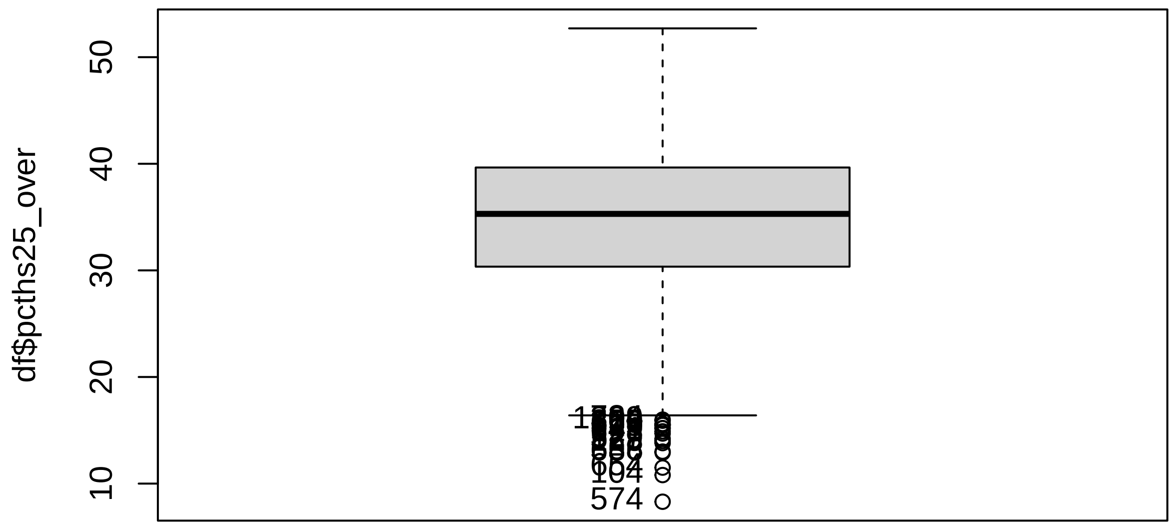
```
## [1] 0
```

```r
Boxplot(df$pcths25_over)
```




```
##  [1] 574 104 654 636 588 128 529 527 628 941
```

```r
length(Boxplot(df$pcths25_over, id = list(n=Inf)))
```




```
## [1] 18
```

```
sevout_pcths25_over = (quantile(df$pcths25_over,0.25)+(3*((quantile(df$pcths25_over,0.75)-quantile(df$pc
length(which(df$pcths25_over > sevout_pcths25_over))
```

```
## [1] 0
```

```
df$f.pcths25_over <- ifelse(df$pcths25_over <= 30.35, 1, ifelse(df$pcths25_over > 30.35 & df$pcths25_ove
df$f.pcths25_over <- factor(df$f.pcths25_over, labels=c("Low25Highsc%","LowMid25Highsc%","HighMid25Highs
table(df$f.pcths25_over)
```

```
##
##    Low25Highsc%  LowMid25Highsc% HighMid25Highsc%    High25Highsc%
##             458              469              446              458
```

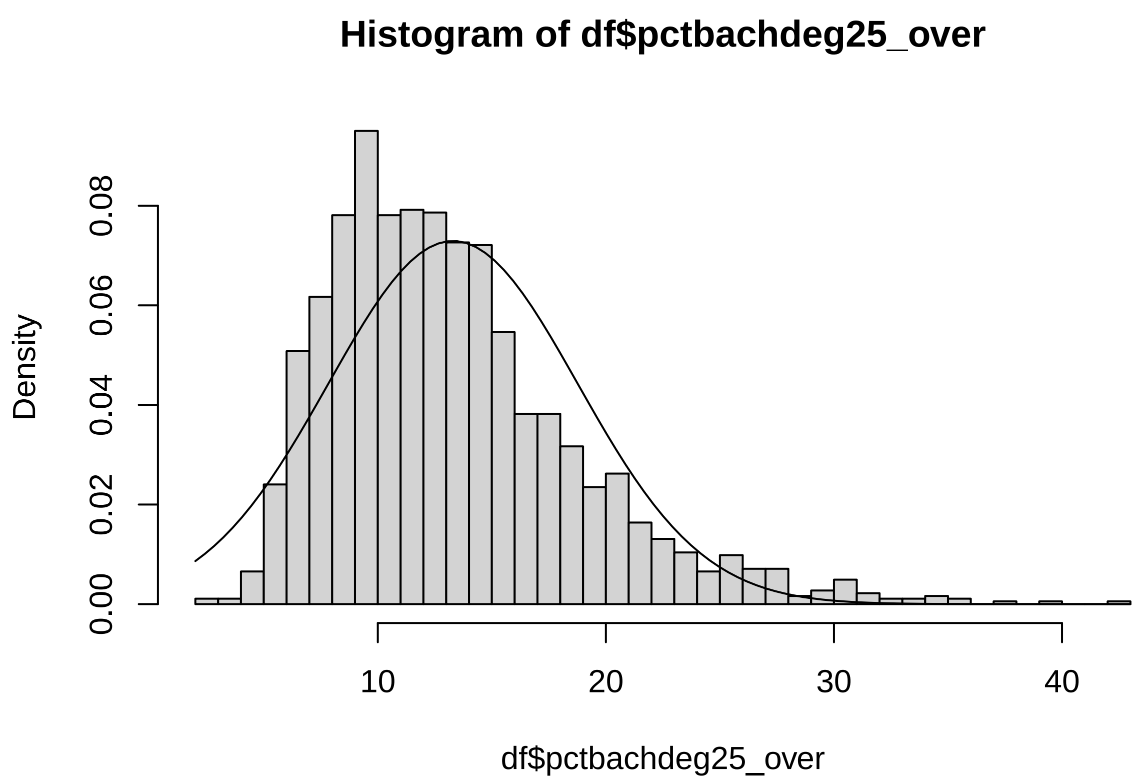
Variable 18 - pctbachdeg25_over

Another continuous ratio variable (related to the previous one) not normally distributed with 0 missing values, 59 outliers (27 of them severe) all on the higher end. We create an additional ordinal factor "f.pctbachdeg25_over".

```
summary(df$pctbachdeg25_over)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     2.5     9.3    12.3    13.3    16.0    42.2
```

```r
hist(df$pctbachdeg25_over, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctbachdeg25_over), sd(df$pctbachdeg25_over)), add = T)
```



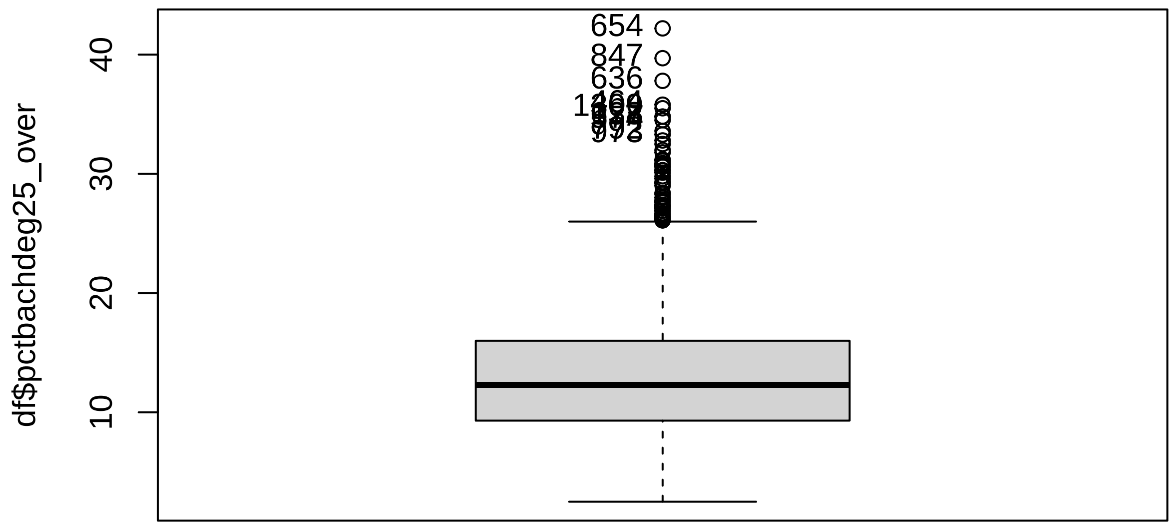

```r
shapiro.test(df$pctbachdeg25_over)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  df$pctbachdeg25_over
## W = 0.92998, p-value < 2.2e-16
```

```r
sum(is.na(df$pctbachdeg25_over))
```

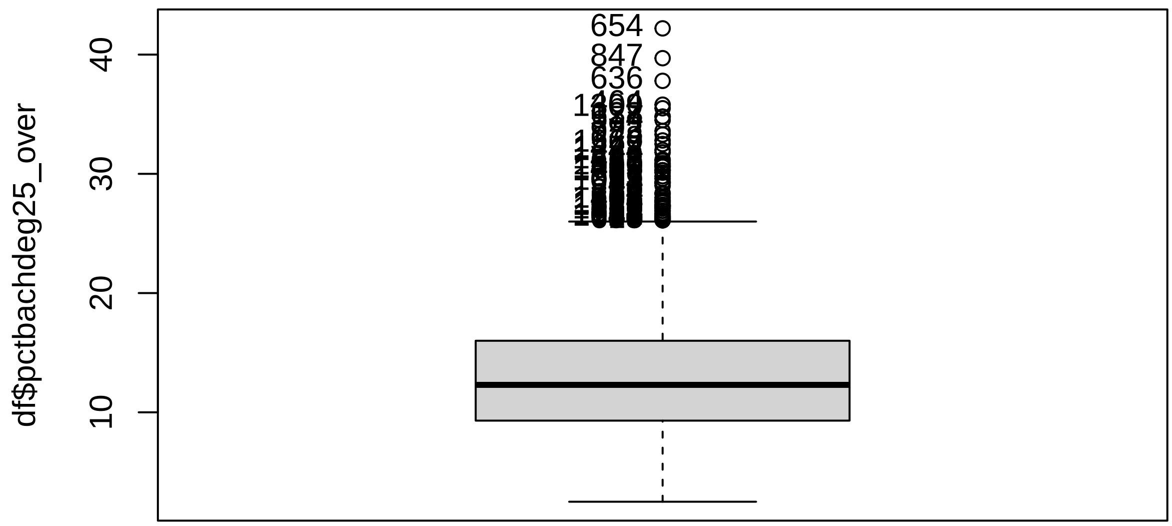
```
## [1] 0
```

```r
Boxplot(df$pctbachdeg25_over)
```




```
## [1]  654  847  636  464 1309  128  637  574  792  973
```

```r
length(Boxplot(df$pctbachdeg25_over, id = list(n=Inf)))
```




```
## [1] 59
```

```
sevout_pctbachdeg25_over = (quantile(df$pctbachdeg25_over,0.25)+(3*((quantile(df$pctbachdeg25_over,0.75
length(which(df$pctbachdeg25_over > sevout_pctbachdeg25_over))
```

```
## [1] 27
```

```
df$f.pctbachdeg25_over <- ifelse(df$pctbachdeg25_over <= 9.3, 1, ifelse(df$pctbachdeg25_over > 9.3 & df$
df$f.pctbachdeg25_over <- factor(df$f.pctbachdeg25_over, labels=c("LowBach%","LowMidBach%","HighMidBach%
table(df$f.pctbachdeg25_over)
```

```
##
##     LowBach%  LowMidBach% HighMidBach%    HighBach%
##          459          458          463          451
```

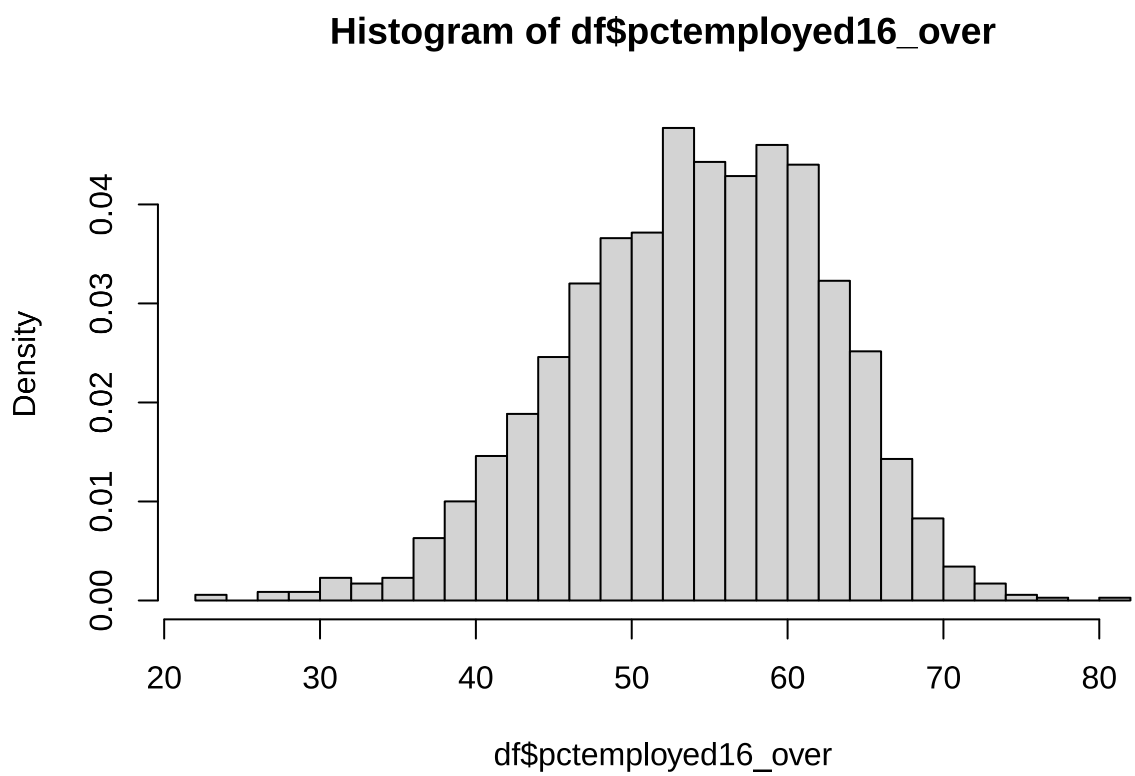
Variable 19 - pctemployed16_over

Another continuous ratio variable not normally distributed with 82 missing values (we will see how to input them later), 11 outliers (none of them severe), all but one on the lower end. We create an additional ordinal factor "f.pctemployed16_over".

```
summary(df$pctemployed16_over)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   23.90   48.60   54.50   54.21   60.30   80.10      82
```

```
hist(df$pctemployed16_over, breaks = 30, freq = F)
```




```
shapiro.test(df$pctemployed16_over)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctemployed16_over
## W = 0.99196, p-value = 3.371e-08
```

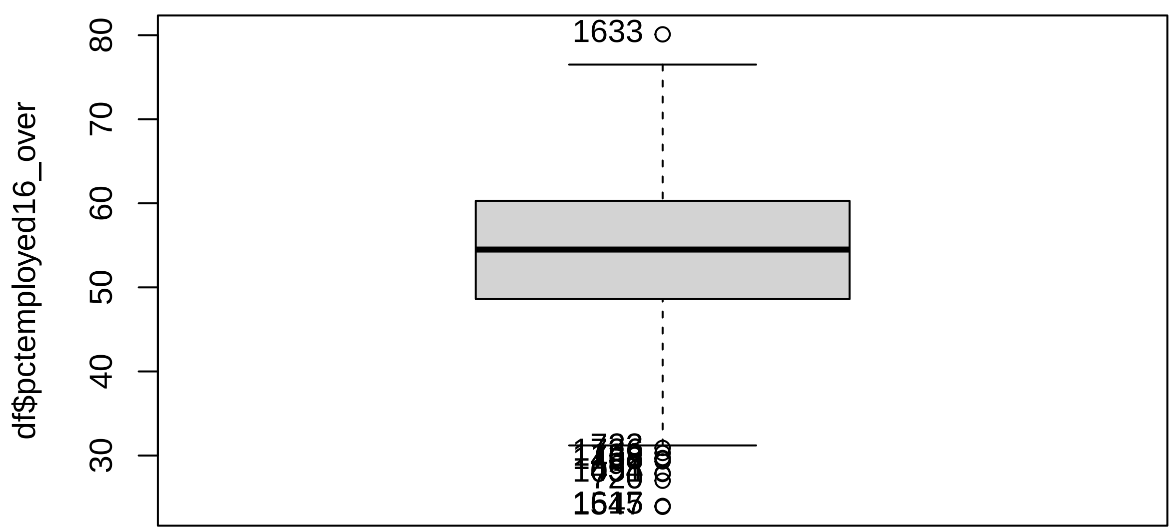
```
sum(is.na(df$pctemployed16_over))
```

```
## [1] 82
```

```
Boxplot(df$pctemployed16_over)
```

```
##  [1]  434  720  723  753 1091 1138 1468 1547 1615 1736 1633
```

```r
length(Boxplot(df$pctemployed16_over, id = list(n=Inf)))
```




```
## [1] 11
```

```r
sevout_pctemployed16_over = (48.60+(3*(60.30-48.60)))
length(which(df$pctemployed16_over > sevout_pctemployed16_over))
```

```
## [1] 0
```

```r
df$f.pctemployed16_over <- ifelse(df$pctemployed16_over <= 48.60, 1, ifelse(df$pctemployed16_over > 48.6
df$f.pctemployed16_over <- factor(df$f.pctemployed16_over, labels=c("LowEmploy%","LowMidEmploy%","HighMi
table(df$f.pctemployed16_over)
```

```
##
##     LowEmploy%  LowMidEmploy% HighMidEmploy%    HighEmploy%
##            442            434            444            429
```

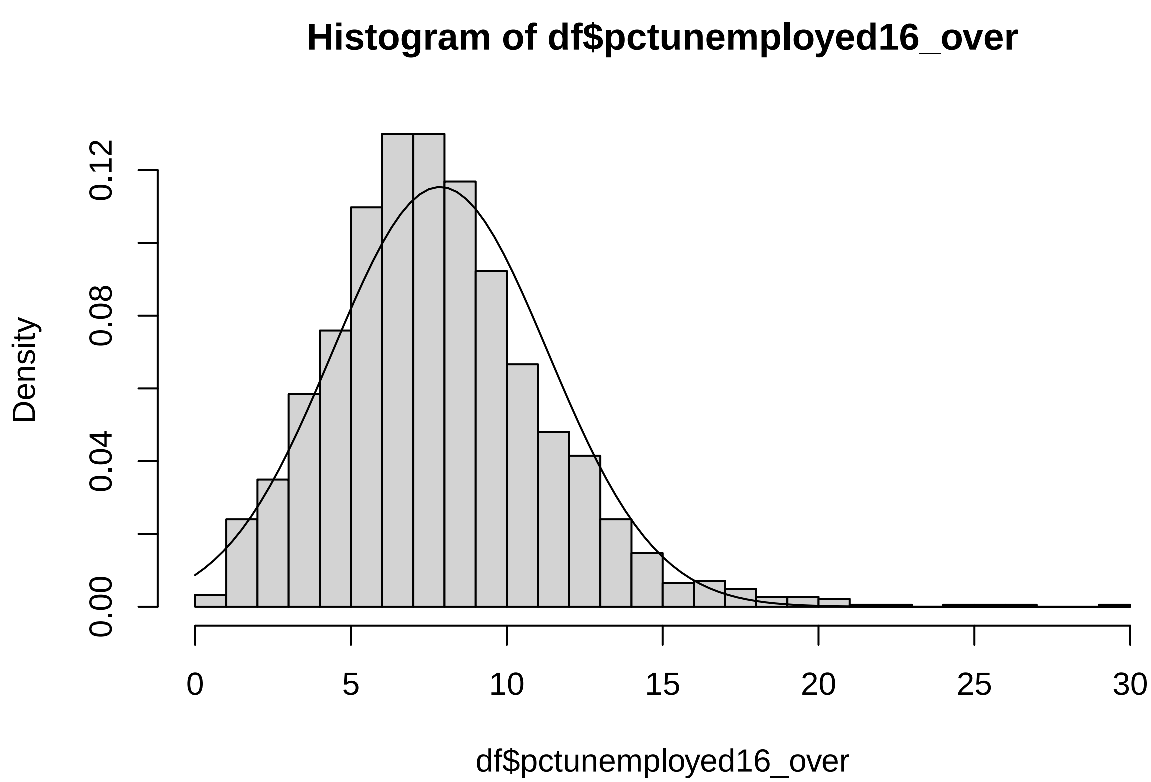
Variable 20 - pctunemployed16_over

One would assume that this variable is 100 minus the previous variable, but looking at some observations this is proven false. It is a continuous ratio variable not normally distributed with 0 missing values, 42 outliers (18 of them severe), all on the higher end. We create an additional ordinal factor “f.pctunemployed16_over”.

```r
summary(df$pctunemployed16_over)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   5.500   7.500   7.861   9.750  29.400
```

```r
hist(df$pctunemployed16_over, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctunemployed16_over), sd(df$pctunemployed16_over)), add = T)
```



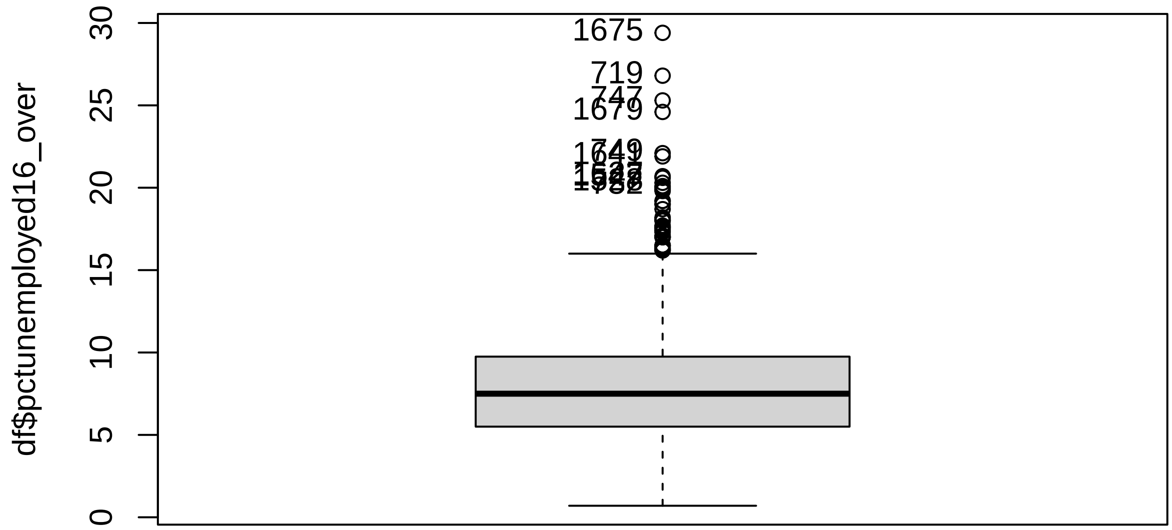

```r
shapiro.test(df$pctunemployed16_over)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctunemployed16_over
## W = 0.9612, p-value < 2.2e-16
```

```r
sum(is.na(df$pctunemployed16_over))
```

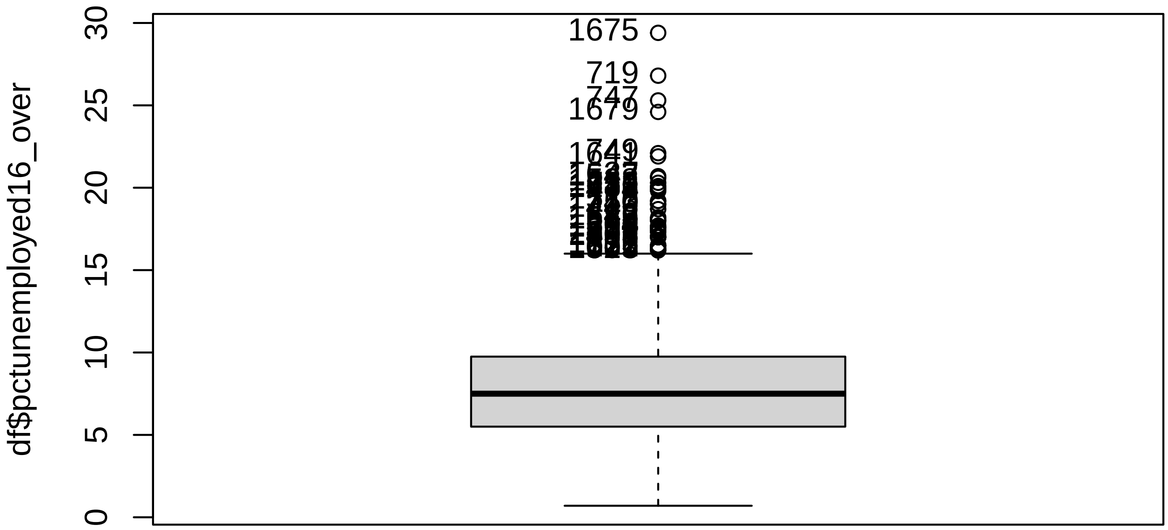
```
## [1] 0
```

```r
Boxplot(df$pctunemployed16_over)
```




```
##  [1] 1675  719  747 1679  749 1641 1622 1547 1528  752
```

```r
length(Boxplot(df$pctunemployed16_over, id = list(n=Inf)))
```




```
## [1] 42
```

```r
sevout_pctunemployed16_over = (quantile(df$pctunemployed16_over,0.25)+(3*((quantile(df$pctunemployed16_o
length(which(df$pctunemployed16_over > sevout_pctunemployed16_over))
```

```
## [1] 18
```

```r
df$f.pctunemployed16_over <- ifelse(df$pctunemployed16_over <= 5.5, 1, ifelse(df$pctunemployed16_over >
df$f.pcuntemployed16_over <- factor(df$f.pctunemployed16_over, labels=c("LowUnEmploy%","LowMidUnEmploy%
table(df$f.pctunemployed16_over)
```

```
##
##   1   2   3   4
## 467 453 453 458
```

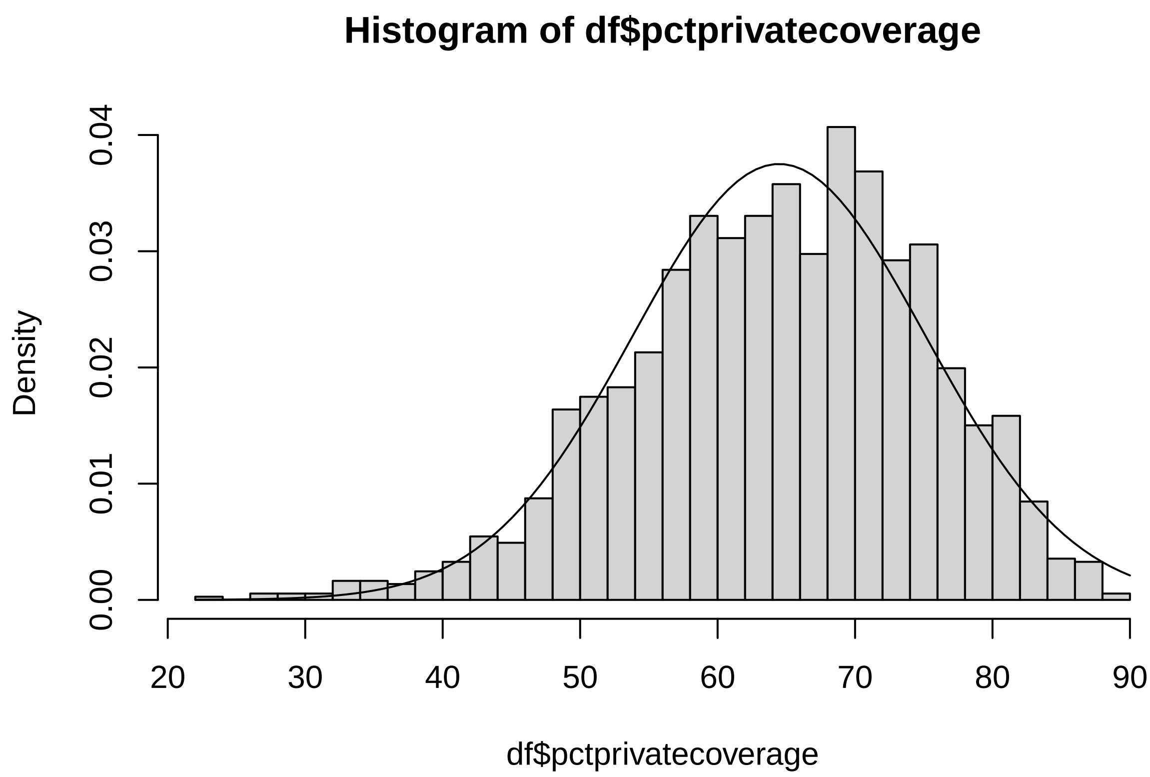
Variable 21 - pctprivatecoverage

Another continuous ratio variable not normally distributed with 0 missing values, 17 outliers (none of them severe) all on the lower end. We create an additional ordinal factor "f.pctprivatecoverage".

```r
summary(df$pctprivatecoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.40   57.50   65.20   64.47   72.10   89.60
```

```r
hist(df$pctprivatecoverage, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctprivatecoverage), sd(df$pctprivatecoverage)), add = T)
```



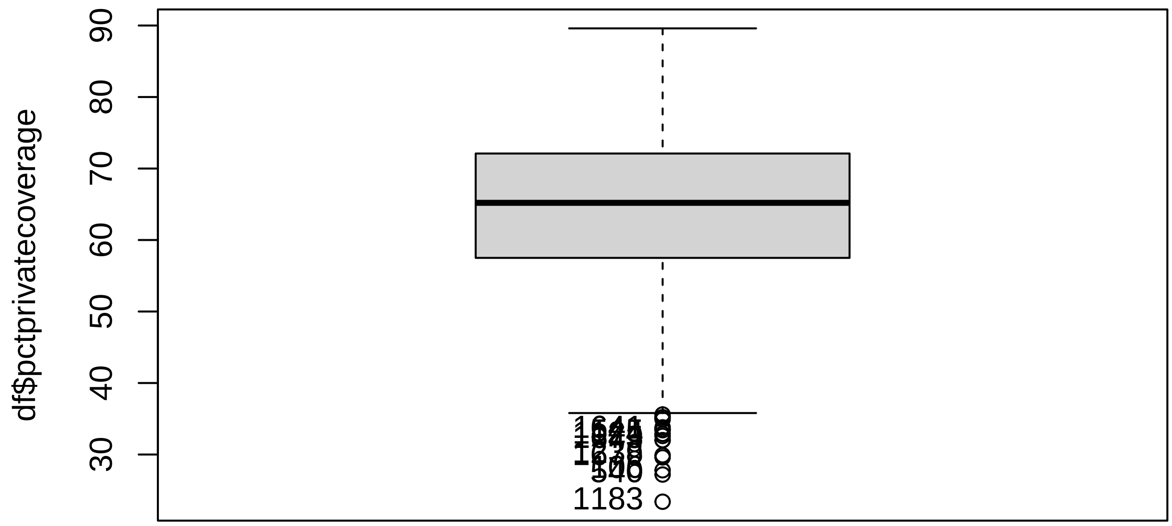

```r
shapiro.test(df$pctprivatecoverage)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctprivatecoverage
## W = 0.98964, p-value = 3.725e-10
```

```r
sum(is.na(df$pctprivatecoverage))
```

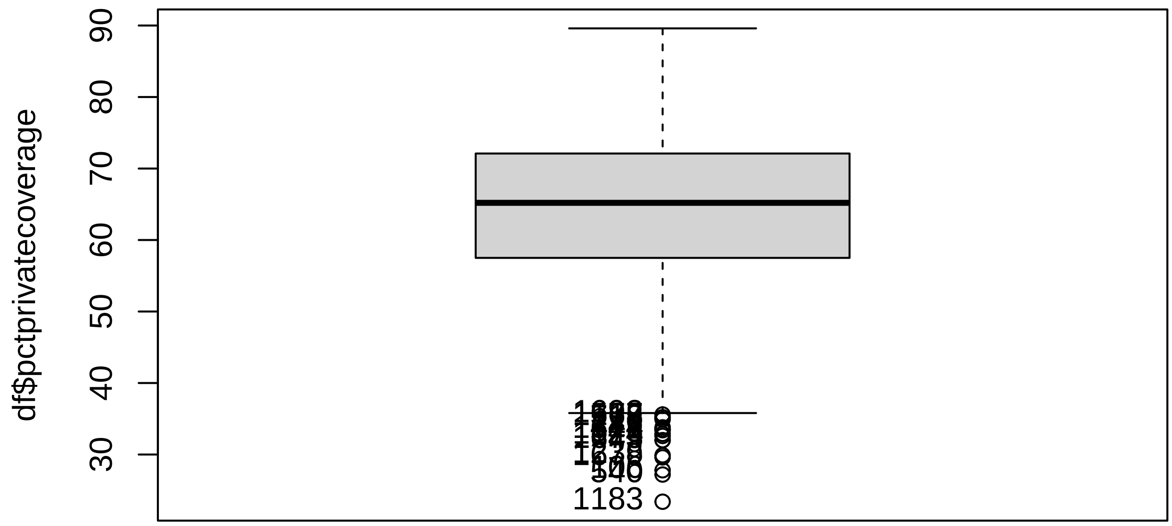
```
## [1] 0
```

```r
Boxplot(df$pctprivatecoverage)
```




```
##  [1] 1183  540  106 1679 1238 1643 1675 1124  545 1641
```

```r
length(Boxplot(df$pctprivatecoverage, id = list(n=Inf)))
```




```
## [1] 17
```

```
sevout_pctprivatecoverage = (quantile(df$pctprivatecoverage,0.25)+(3*((quantile(df$pctprivatecoverage,0
length(which(df$pctprivatecoverage > sevout_pctprivatecoverage))
```

```
## [1] 0
```

```
df$f.pctprivatecoverage <- ifelse(df$pctprivatecoverage <= 57.50, 1, ifelse(df$pctprivatecoverage > 57.5
df$f.pctprivatecoverage <- factor(df$f.pctprivatecoverage, labels=c("LowPrivate%","LowMidPrivate%","High
table(df$f.pctprivatecoverage)
```

```
##
##     LowPrivate%  LowMidPrivate% HighMidPrivate%    HighPrivate%
##             460             464             451             456
```

Variable 22 - pctprivatecoveragealone

This is a continuous ratio variable very closely related with the previous variable. It also has 356 missing values, which amounts to almost 20% of the observations. Since the number of missing values is high and it doesn't add much to our data (it has a correlation of 0.93 with the previous variable) we will delete it.

```
summary(df$pctprivatecoveragealone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   16.80   41.50   49.00   48.65   55.50   78.90     356
```

```r
sum(is.na(df$pctprivatecoveragealone))
```

```
## [1] 356
```

```r
356/1831*100
```

```
## [1] 19.44293
```

```r
cor.test(df$pctprivatecoverage, df$pctprivatecoveragealone)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$pctprivatecoverage and df$pctprivatecoveragealone
## t = 98.883, df = 1473, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9252270 0.9386221
## sample estimates:
##       cor
## 0.9322432
```

```r
df <- subset(df, select = -pctprivatecoveragealone)
```

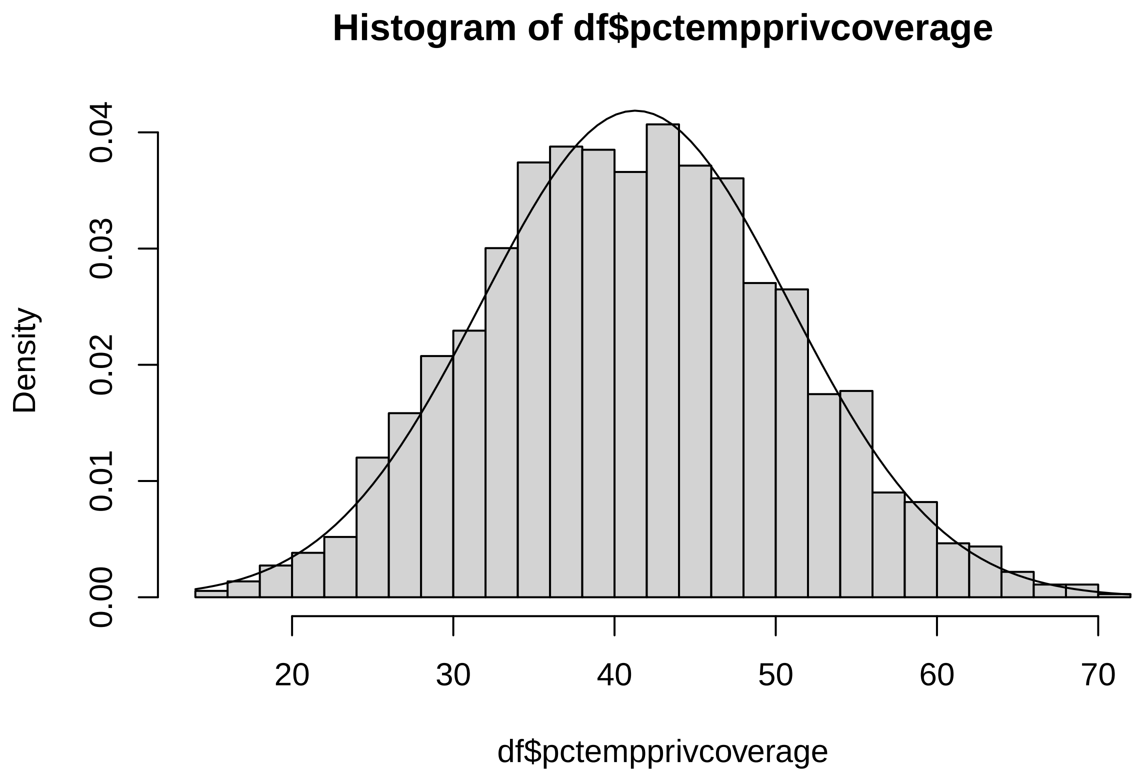
Variable 22 - pctempprivcoverage

Another continuous ratio variable normally distributed (if we pick a 99% significance level for the shapiro test) with 0 missing values, 7 outliers (none of them severe) all on the higher end but one. We create an additional ordinal factor "f.pctempprivcoverage".

```r
summary(df$pctempprivcoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.30   34.60   41.10   41.29   47.70   70.20
```

```r
hist(df$pctempprivcoverage, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctempprivcoverage), sd(df$pctempprivcoverage)), add = T)
```




```r
shapiro.test(df$pctempprivcoverage)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctempprivcoverage
## W = 0.99807, p-value = 0.02861
```

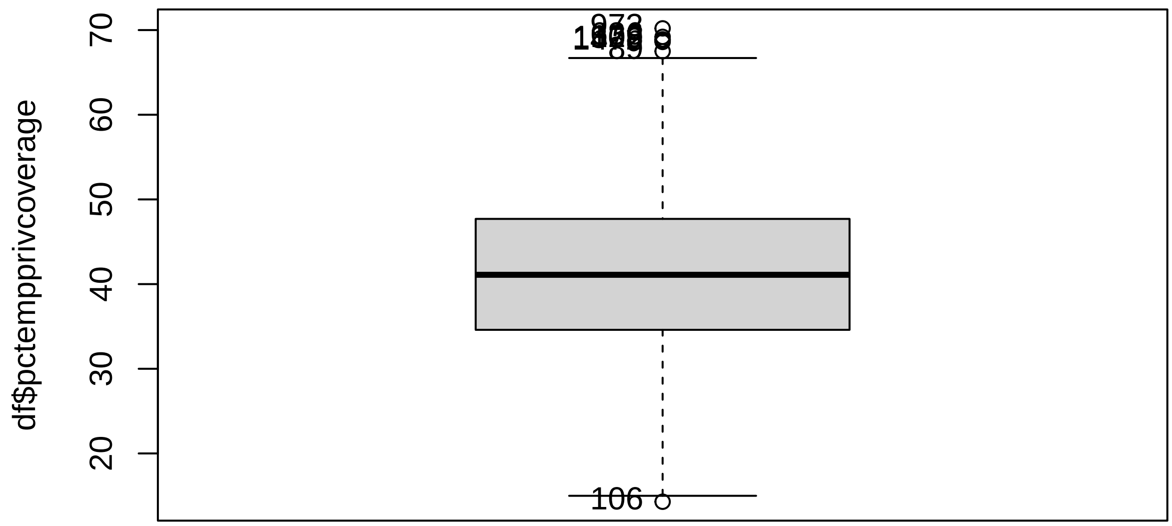
```r
sum(is.na(df$pctempprivcoverage))
```

```
## [1] 0
```

```r
Boxplot(df$pctempprivcoverage)
```

```
## [1]  106   89  128  636  973 1309 1472
```

```r
length(Boxplot(df$pctempprivcoverage, id = list(n=Inf)))
```




```
## [1] 7
```

```r
sevout_pctempprivcoverage = (quantile(df$pctempprivcoverage,0.25)+(3*((quantile(df$pctempprivcoverage,0
length(which(df$pctempprivcoverage > sevout_pctempprivcoverage))
```

```
## [1] 0
```

```r
df$f.pctempprivcoverage <- ifelse(df$pctempprivcoverage <= 34.60, 1, ifelse(df$pctempprivcoverage > 34.
df$f.pctempprivcoverage <- factor(df$f.pctempprivcoverage, labels=c("LowEmployeeHealth%","LowMidEmployee
table(df$f.pctempprivcoverage)
```

```
##
##     LowEmployeeHealth%  LowMidEmployeeHealth% HighMidEmployeeHealth%
##                    465                    454                    456
##    HighEmployeeHealth%
##                    456
```

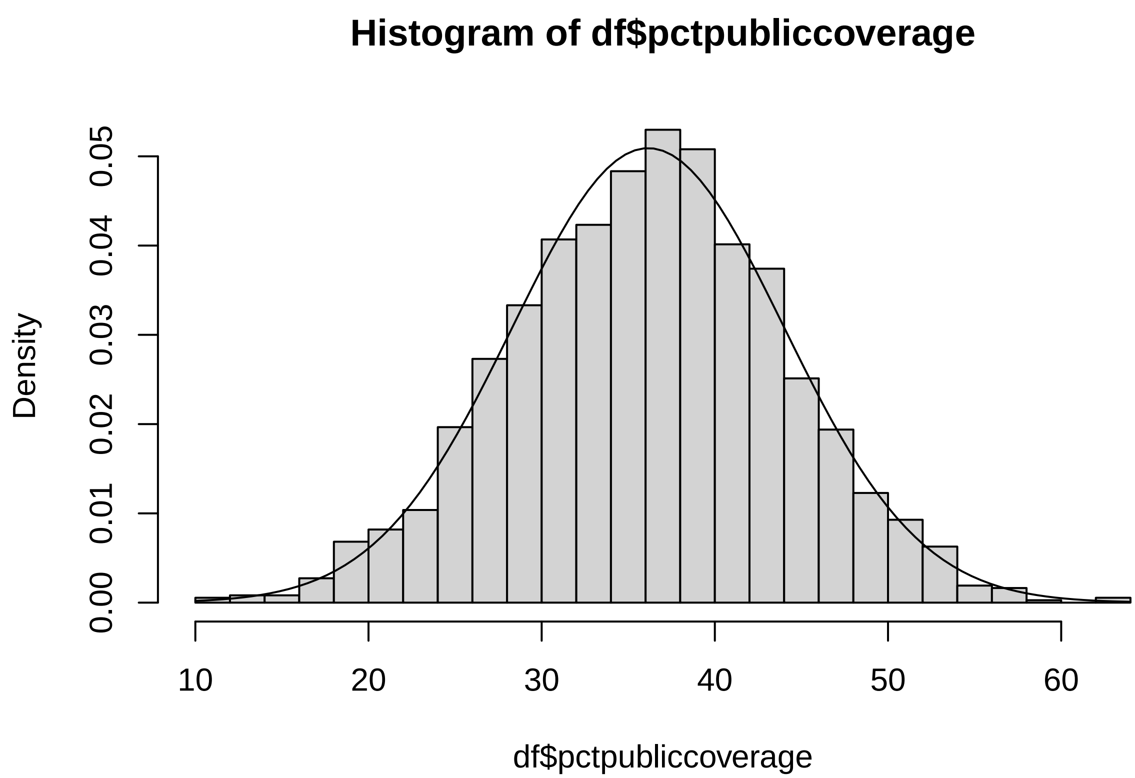
Variable 23 - pctpubliccoverage

Another continuous ratio variable normally distributed with 0 missing values, 13 outliers (1 of them severe) on both ends of the spectrum. We create an additional ordinal factor "f.pctpubliccoverage".

```r
summary(df$pctpubliccoverage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.20   30.90   36.30   36.15   41.40   62.70
```

```r
hist(df$pctpubliccoverage, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctpubliccoverage), sd(df$pctpubliccoverage)), add = T)
```




```r
shapiro.test(df$pctpubliccoverage)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctpubliccoverage
## W = 0.99947, p-value = 0.9186
```

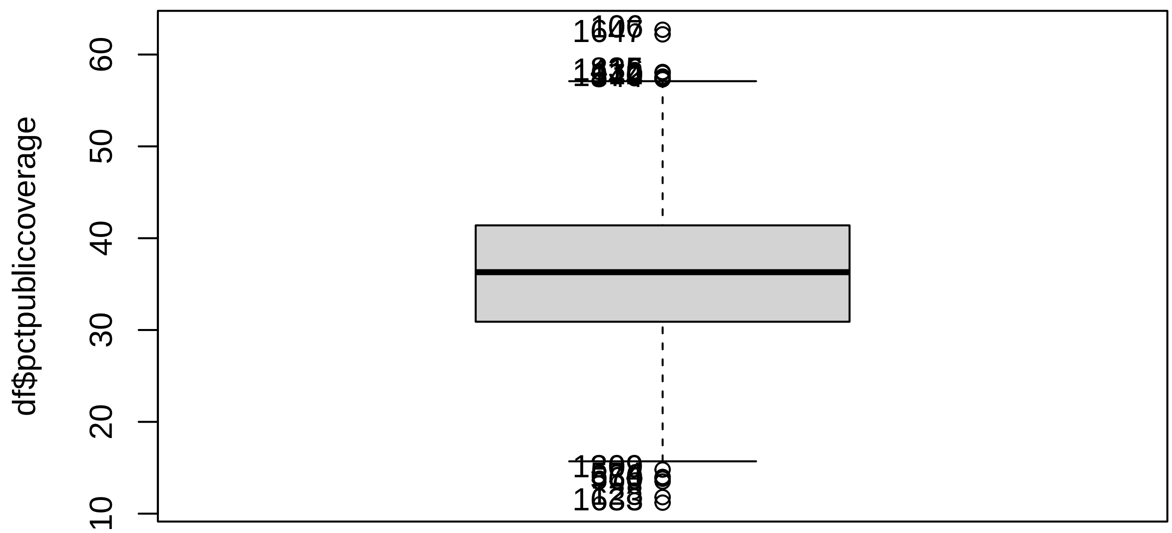
```r
sum(is.na(df$pctpubliccoverage))
```

```
## [1] 0
```

```r
Boxplot(df$pctpubliccoverage)
```

```
##  [1]  128  560  574  636 1309 1633  106  112  835  844 1416 1570 1647
```

```r
length(Boxplot(df$pctpubliccoverage, id = list(n=Inf)))
```



```
## [1] 13
```

```r
sevout_pctpubliccoverage = (quantile(df$pctpubliccoverage,0.25)+(3*((quantile(df$pctpubliccoverage,0.75)
length(which(df$pctpubliccoverage > sevout_pctpubliccoverage))
```

```
## [1] 1
```

```r
df$f.pctpubliccoverage <- ifelse(df$pctpubliccoverage <= 30.90, 1, ifelse(df$pctpubliccoverage > 30.90 &
df$f.pctpubliccoverage <- factor(df$f.pctpubliccoverage, labels=c("LowGovHealth%","LowMidGovHealth%","Hi
table(df$f.pctpubliccoverage)
```

```
##
##     LowGovHealth%  LowMidGovHealth% HighMidGovHealth%    HighGovHealth%
##               463               459               454               455
```

Variable 24 - pctpubliccoveragealone

Another continuous ratio variable related to the previous variable (this time with no NAs and not as closely correlated as variables 21 and 22, cor=0.87, so we will keep de variable for now) not normally distributed with 0 missing values, 21 outliers (7 of them severe) on the higher end (except one). We create an additional ordinal factor "f.pctpubliccoveragealone".

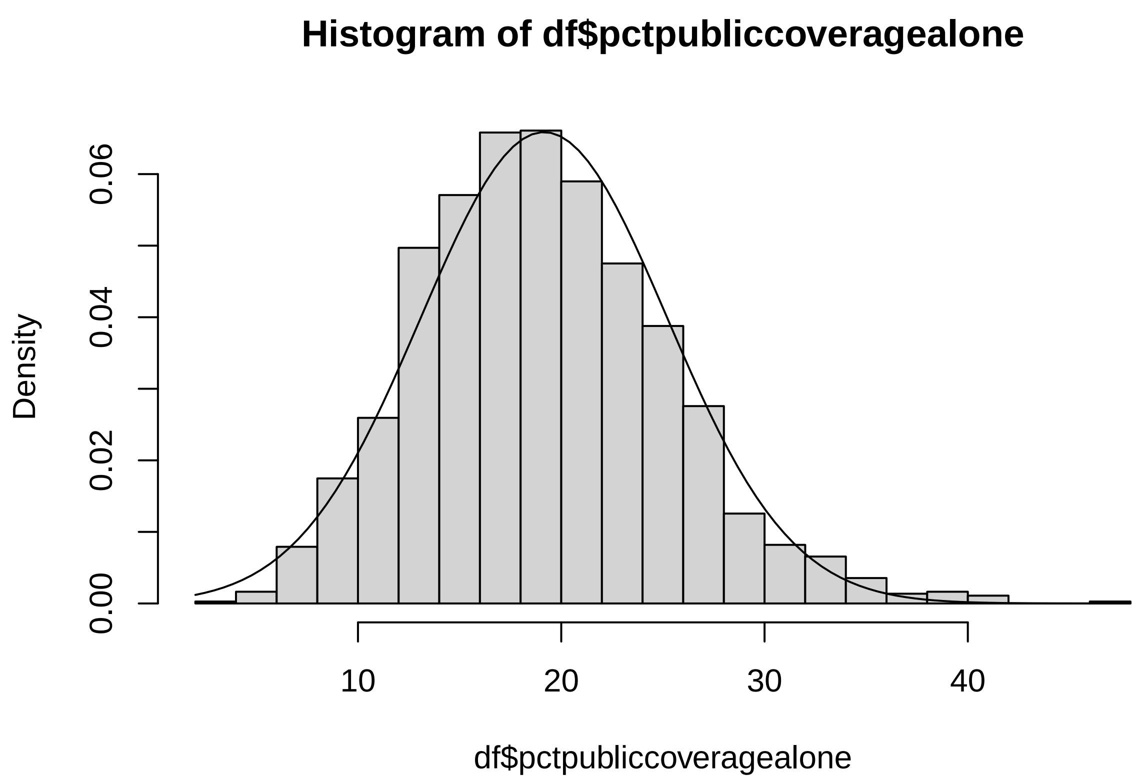
```r
summary(df$pctpubliccoveragealone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.60   14.90   18.70   19.15   23.00   46.60
```

```r
cor.test(df$pctpubliccoverage, df$pctpubliccoveragealone)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$pctpubliccoverage and df$pctpubliccoveragealone
## t = 74.592, df = 1829, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8557240 0.8784263
## sample estimates:
##       cor
## 0.8675263
```

```r
hist(df$pctpubliccoveragealone, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctpubliccoveragealone), sd(df$pctpubliccoveragealone)), add = T)
```



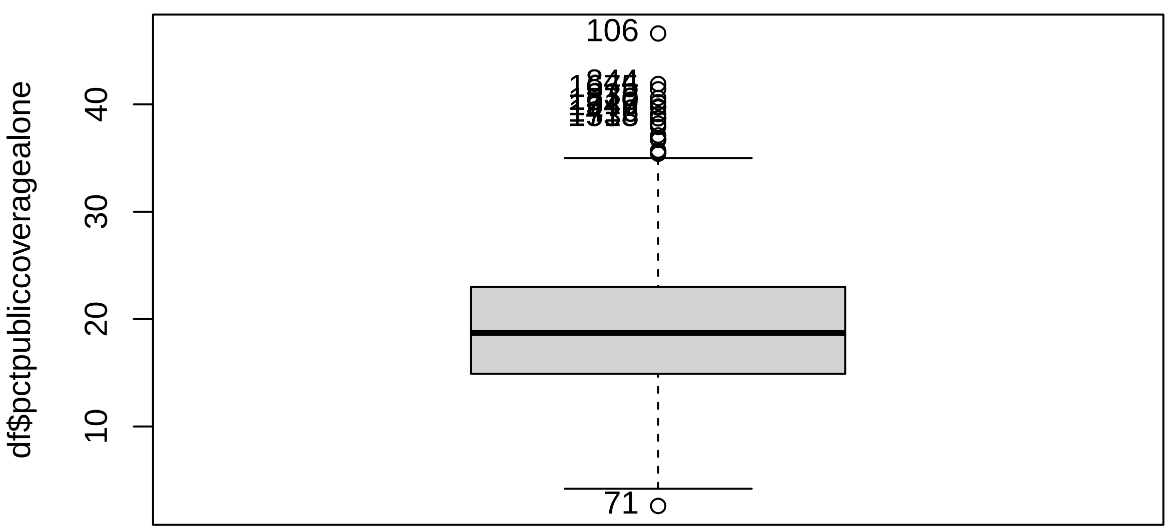
```r
shapiro.test(df$pctpubliccoveragealone)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctpubliccoveragealone
## W = 0.98784, p-value = 2.648e-11
```

```r
sum(is.na(df$pctpubliccoveragealone))
```

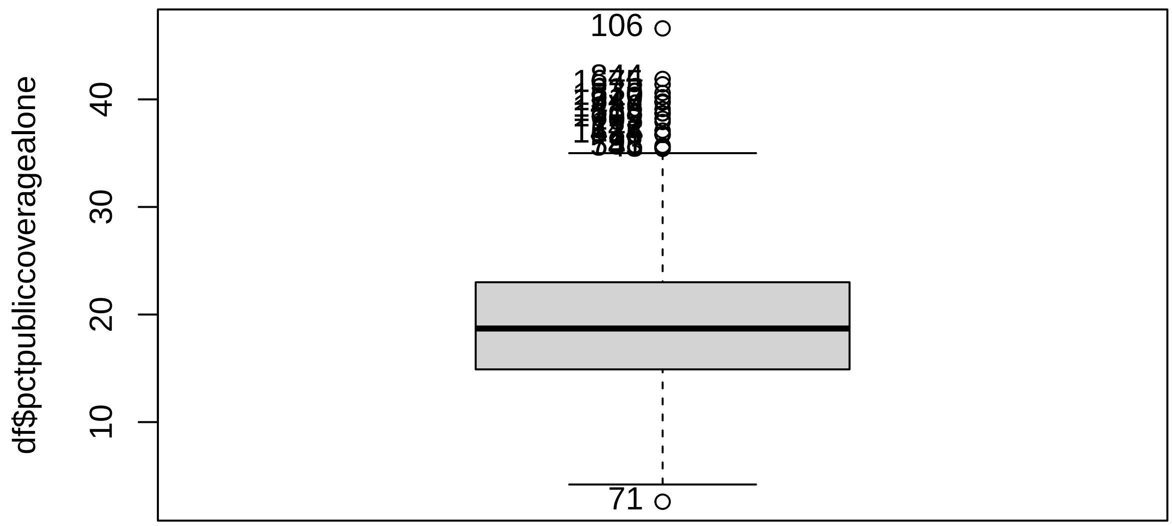
```
## [1] 0
```

```r
Boxplot(df$pctpubliccoveragealone)
```




```
##  [1]   71  106  844 1675  835 1570  719 1547 1416  718 1533
```

```r
length(Boxplot(df$pctpubliccoveragealone, id = list(n=Inf)))
```




```
## [1] 21
```

```
sevout_pctpubliccoveragealone = (quantile(df$pctpubliccoveragealone,0.25)+(3*((quantile(df$pctpubliccove
length(which(df$pctpubliccoveragealone > sevout_pctpubliccoveragealone))
```

```
## [1] 7
```

```
df$f.pctpubliccoveragealone <- ifelse(df$pctpubliccoveragealone <= 14.90, 1, ifelse(df$pctpubliccoverag
df$f.pctpubliccoveragealone <- factor(df$f.pctpubliccoveragealone, labels=c("LowGovHealthAlone%","LowMi
table(df$f.pctpubliccoveragealone)
```

```
##
##     LowGovHealthAlone%  LowMidGovHealthAlone% HighMidGovHealthAlone%
##                    463                    463                    455
##    HighGovHealthAlone%
##                    450
```

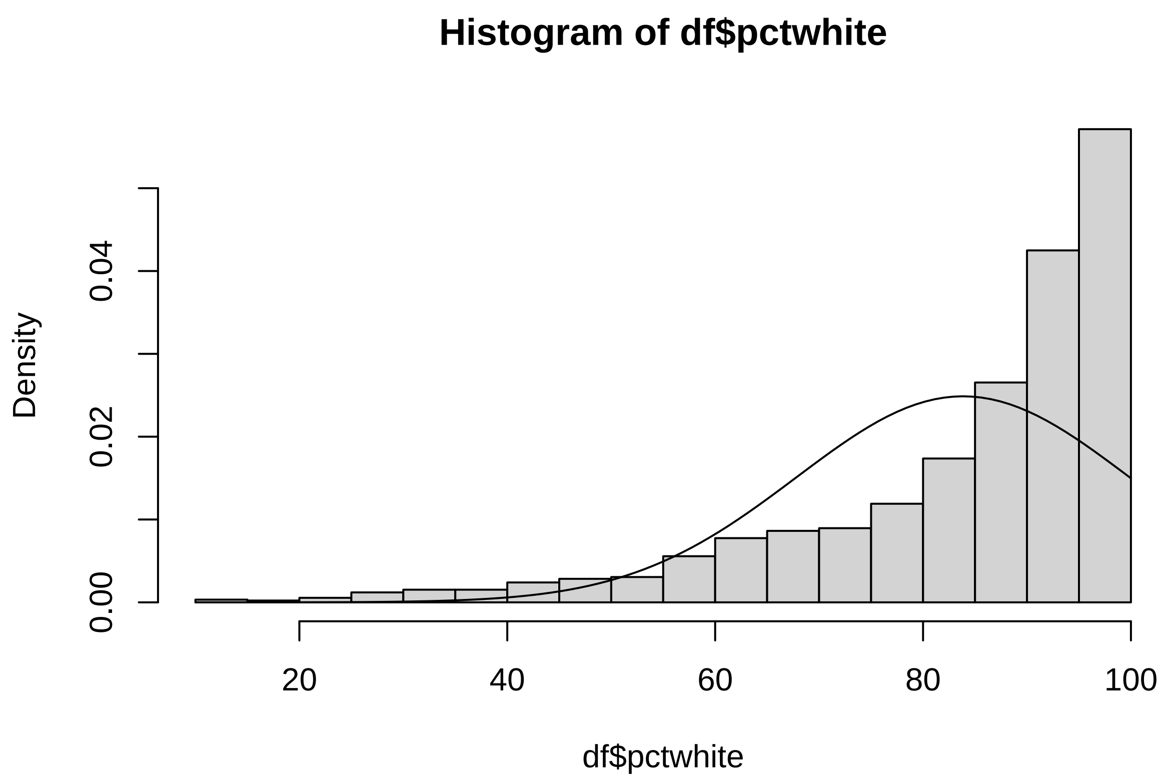
Variable 25 - pctwhite

Another continuous ratio variable clearly not normally distributed with 0 missing values, 97 outliers (none of them severe) all on the low end of the spectrum. We create an additional ordinal factor "f.pctwhite".

```r
summary(df$pctwhite)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.27   77.31   89.90   83.85   95.57   99.69
```

```r
hist(df$pctwhite, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctwhite), sd(df$pctwhite)), add = T)
```



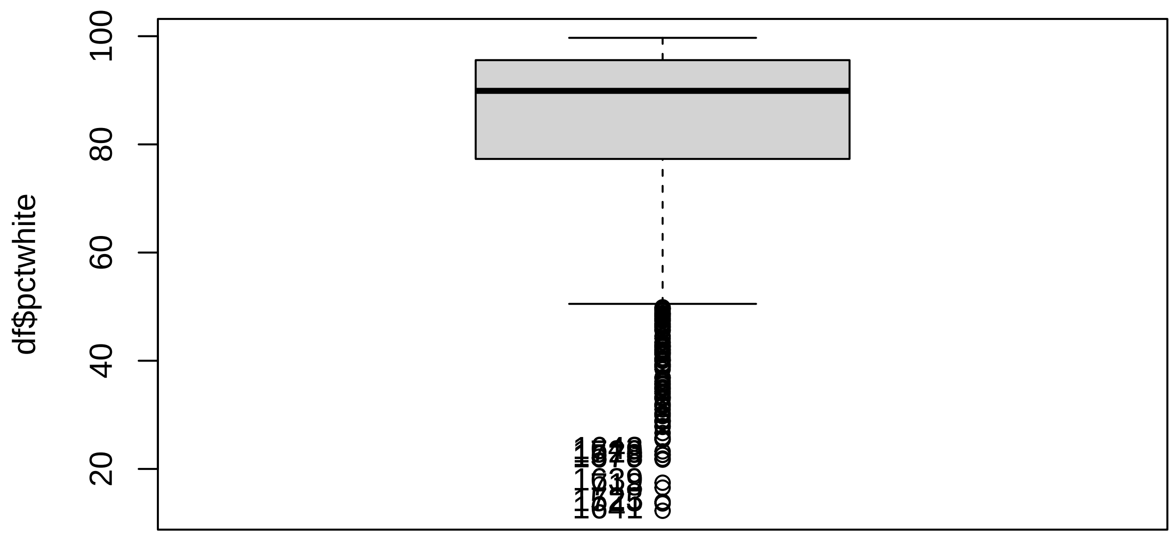

```r
shapiro.test(df$pctwhite)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctwhite
## W = 0.80758, p-value < 2.2e-16
```

```r
sum(is.na(df$pctwhite))
```

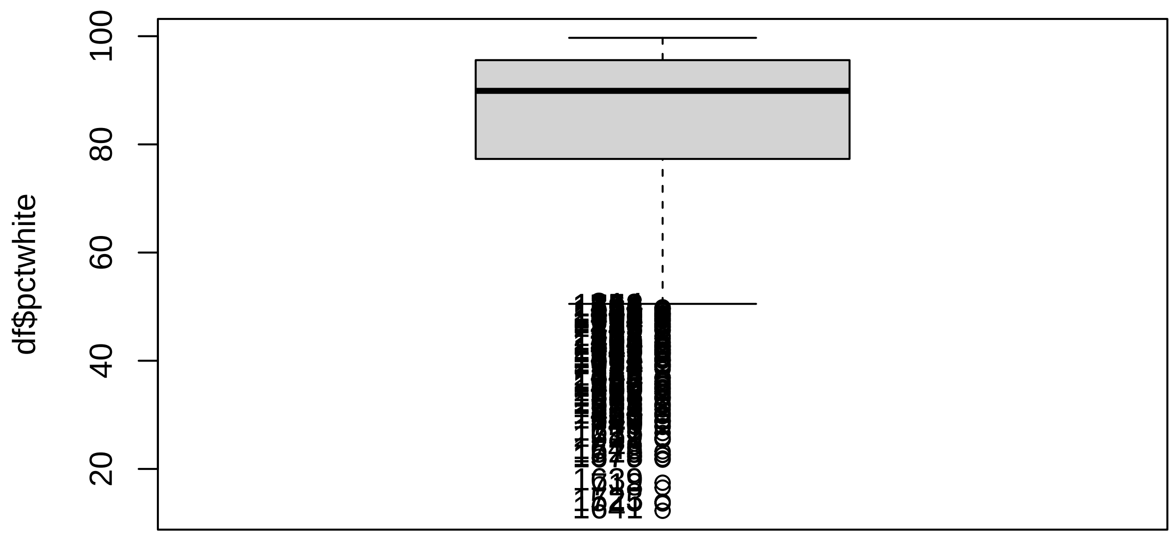
```
## [1] 0
```

```r
Boxplot(df$pctwhite)
```




```
## [1] 1641 1525 723 718 1639 1679 1578 719 1528 1643
```

```
length(Boxplot(df$pctwhite, id = list(n=Inf)))
```




```
## [1] 97
```

```r
sevout_pctwhite = (quantile(df$pctwhite,0.25)+(3*((quantile(df$pctwhite,0.75)-quantile(df$pctwhite,0.25
length(which(df$pctwhite > sevout_pctwhite))
```

```
## [1] 0
```

```r
df$f.pctwhite <- ifelse(df$pctwhite <= 77.31, 1, ifelse(df$pctwhite > 77.31 & df$pctwhite <= 89.90, 2, 
df$f.pctwhite <- factor(df$f.pctwhite, labels=c("LowWhite%","LowMidWhite%","HighMidWhite%","HighWhite%"
table(df$f.pctwhite)
```

```
##
##     LowWhite%  LowMidWhite% HighMidWhite%    HighWhite%
##           458           459           456           458
```

Variable 26 - pctblack

Really similar to the previous variable, with a correlation of 0.84. It is another continuous ratio variable clearly not normally distributed with 0 missing values, 224 outliers (168 of them severe) all on the high end of the spectrum. We create an additional ordinal factor "f.pctblack".

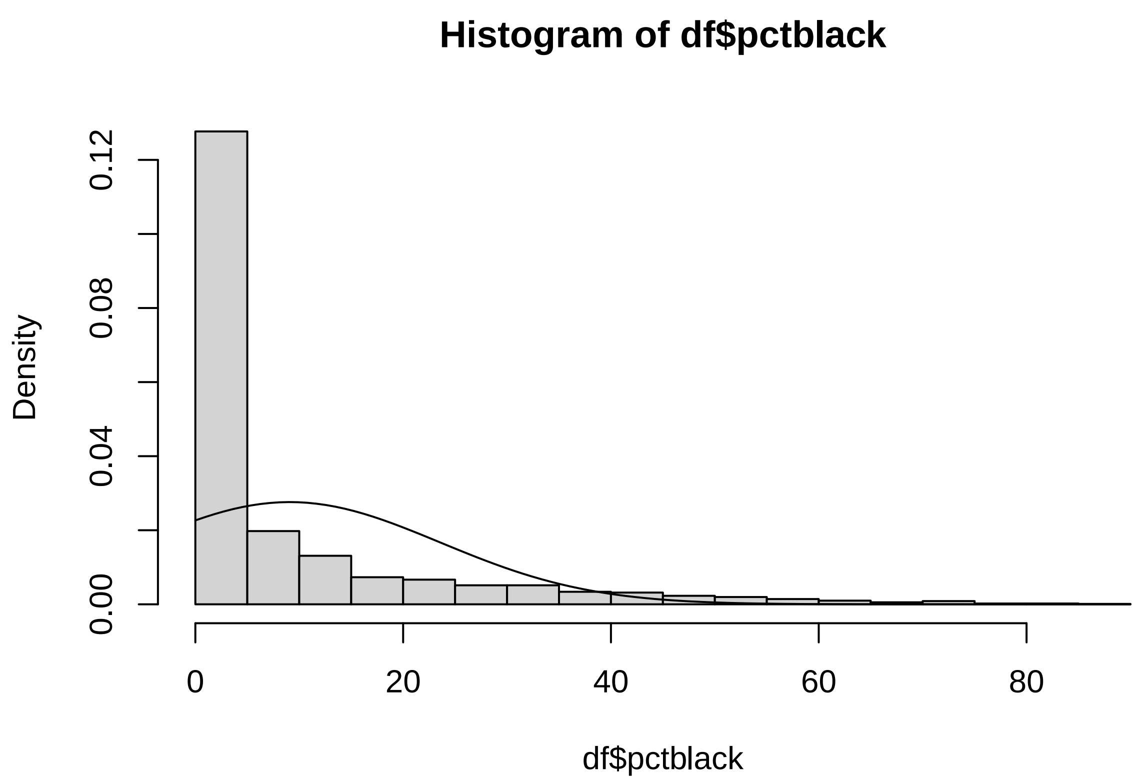
```r
summary(df$pctblack)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.648   2.323   9.082  10.867  85.948
```

```
cor.test(df$pctwhite, df$pctblack)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$pctwhite and df$pctblack
## t = -67.439, df = 1829, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8571535 -0.8308366
## sample estimates:
##        cor
## -0.8445041
```

```
hist(df$pctblack, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctblack), sd(df$pctblack)), add = T)
```




```
shapiro.test(df$pctblack)
```

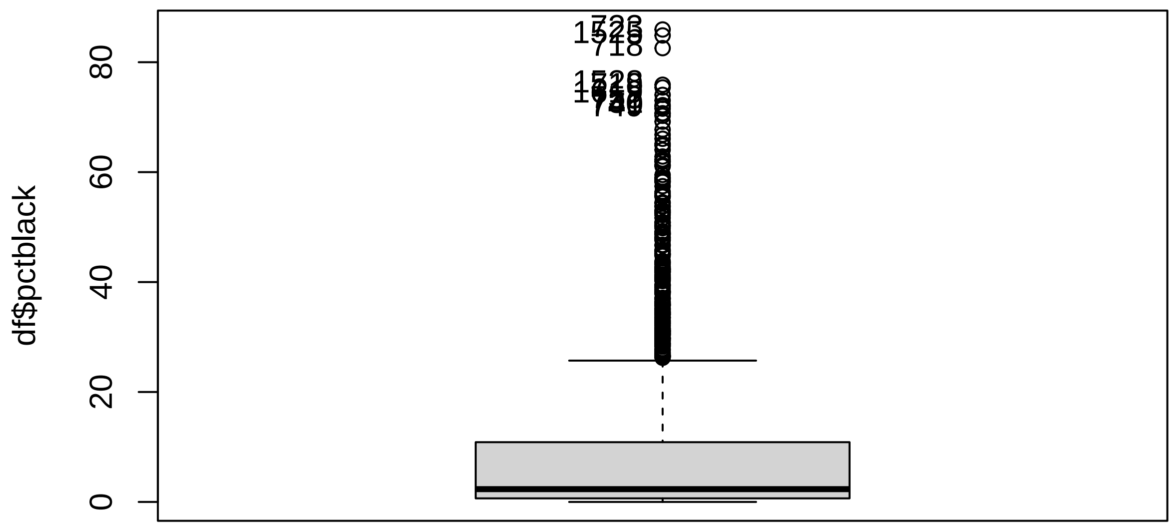
```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  df$pctblack
## W = 0.65926, p-value < 2.2e-16
```

```r
sum(is.na(df$pctblack))
```

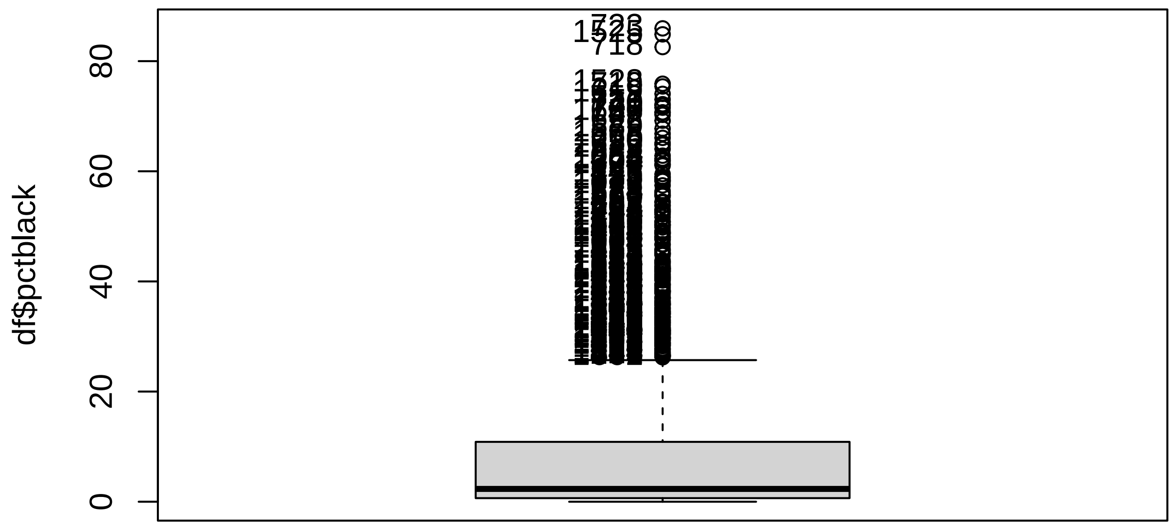
```
## [1] 0
```

```r
Boxplot(df$pctblack)
```



```
##  [1]  723 1525  718 1528  719 1619  752  731  740  749
```

```r
length(Boxplot(df$pctblack, id = list(n=Inf)))
```



```
## [1] 224
```

```
sevout_pctblack = (quantile(df$pctblack,0.25)+(3*((quantile(df$pctblack,0.75)-quantile(df$pctblack,0.25)
length(which(df$pctblack > sevout_pctblack))
```

```
## [1] 168
```

```
df$f.pctblack <- ifelse(df$pctblack <= 0.648, 1, ifelse(df$pctblack > 0.648 & df$pctblack <= 2.323, 2, 
df$f.pctblack <- factor(df$f.pctblack, labels=c("LowBlack%","LowMidBlack%","HighMidBlack%","HighBlack%"
table(df$f.pctblack)
```

```
##
##      LowBlack%  LowMidBlack% HighMidBlack%    HighBlack%
##            458           459           456           458
```

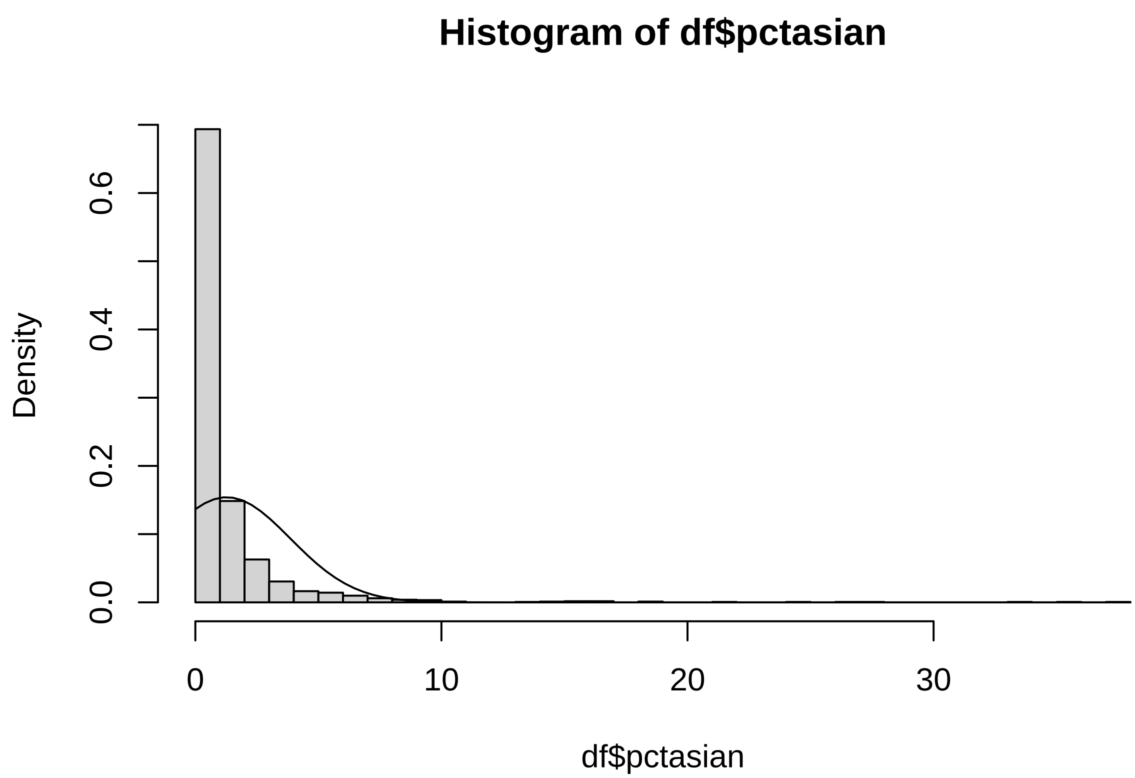
Variable 27 - pctasian

Also related to the previous 2 variables. It is a continuous ratio variable clearly not normally distributed with 0 missing values, 198 outliers (156 of them severe, and looking at the boxplot some of them really far, probably asian ghetto counties) all on the high end of the spectrum. We create an additional ordinal factor "f.pctasian".

```r
summary(df$pctasian)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2582  0.5495  1.2743  1.2515 37.1569
```

```r
hist(df$pctasian, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctasian), sd(df$pctasian)), add = T)
```



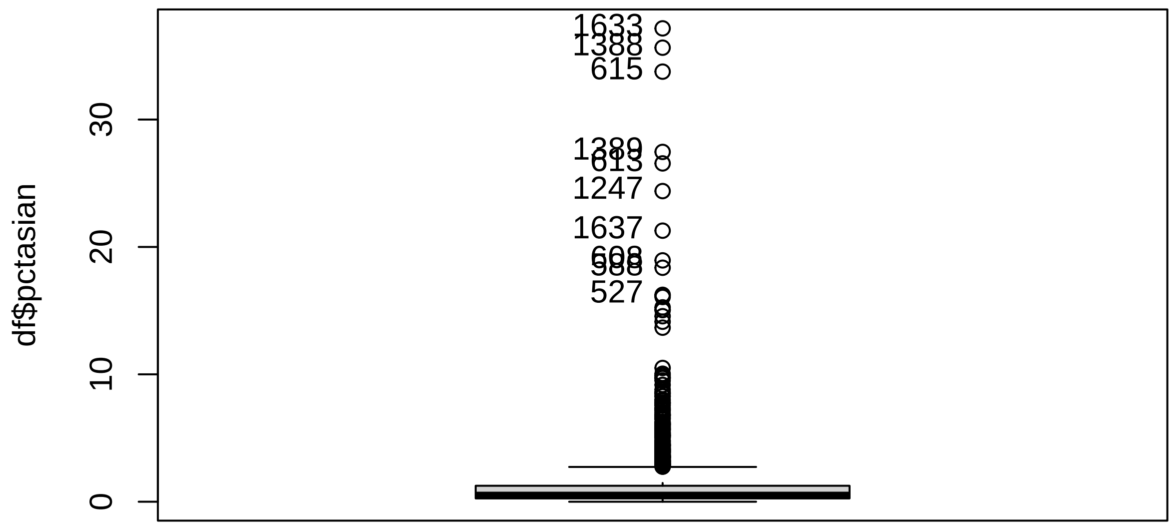

```r
shapiro.test(df$pctasian)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctasian
## W = 0.41908, p-value < 2.2e-16
```

```r
sum(is.na(df$pctasian))
```

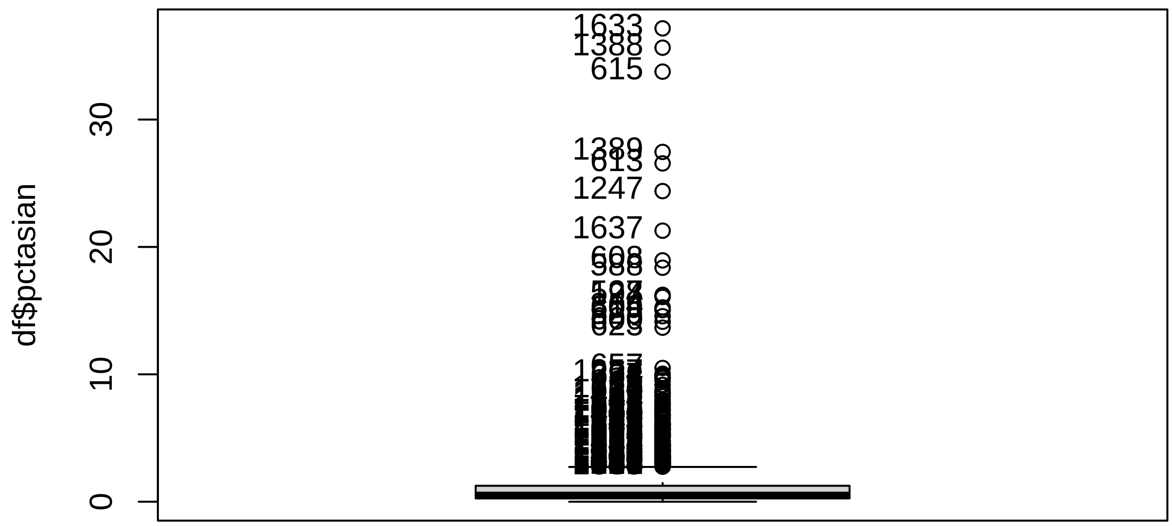
```
## [1] 0
```

```r
Boxplot(df$pctasian)
```




```
## [1] 1633 1388  615 1389  613 1247 1637  608  588  527
```

```r
length(Boxplot(df$pctasian, id = list(n=Inf)))
```




```
## [1] 198
```

```
sevout_pctasian = (quantile(df$pctasian,0.25)+(3*((quantile(df$pctasian,0.75)-quantile(df$pctasian,0.25
length(which(df$pctasian > sevout_pctasian))
```

```
## [1] 156
```

```
df$f.pctasian <- ifelse(df$pctasian <= 0.2582, 1, ifelse(df$pctasian > 0.2582 & df$pctasian <= 0.5495, 
df$f.pctasian <- factor(df$f.pctasian, labels=c("LowAsian%","LowMidAsian%","HighMidAsian%","HighAsian%"
table(df$f.pctasian)
```

```
##
##     LowAsian%  LowMidAsian% HighMidAsian%    HighAsian%
##           458           457           458           458
```

Variable 28 - pctotherrace

This variable should be 100 minus the sum of the three previous variables but looking at a sample of observations it is clearly not, and also if we check for multicollinearity using VIF, since the values are lower than 5 we can use the rule of thumb to say that there is not a severe multicollinearity so we will keep the variable for now (if it was always equal to 100 we would erase it since it wouldn't add any new info). The variable is a continuous ratio variable clearly not normally distributed with 0 missing values, 181 outliers (148 of them severe, and looking at the boxplot some of them really far, probably asian ghetto counties) all on the high end of the spectrum. We create an additional ordinal factor "f.pctotherrace".

```
summary(df$pctotherrace)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.2867   0.7826   2.0031   2.1066  41.9303
```

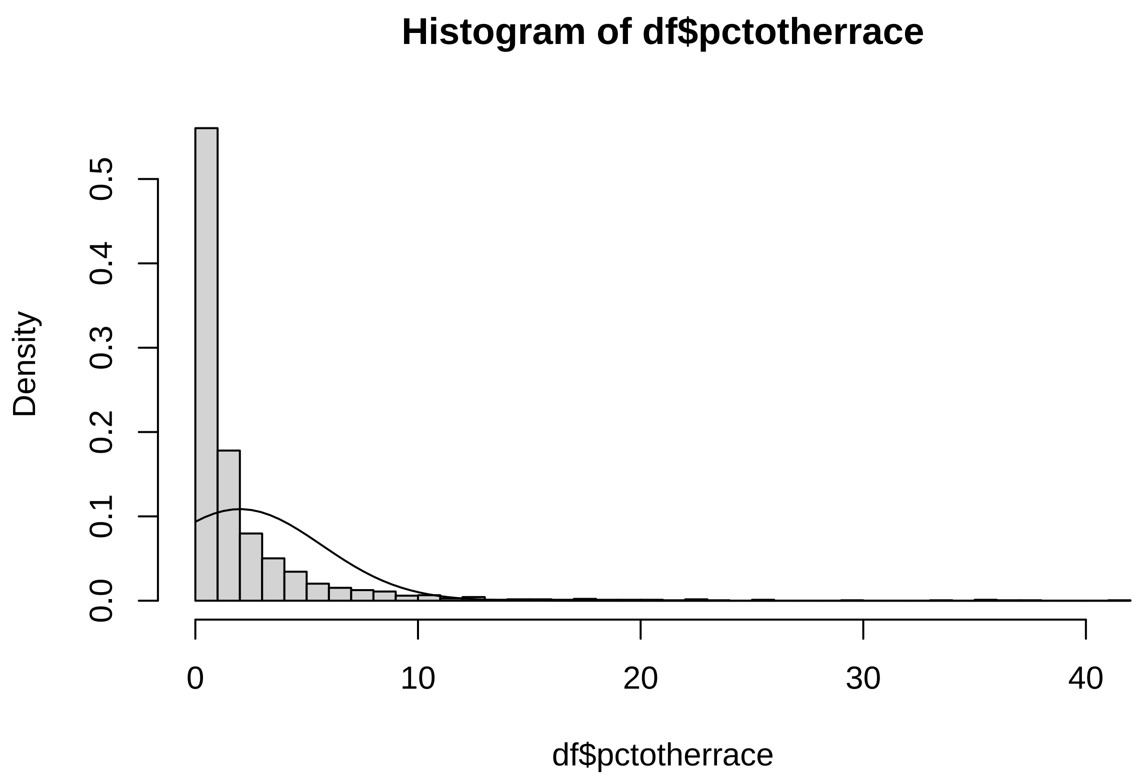
```
model <- lm(pctotherrace ~ pctwhite + pctblack + pctasian, data=df)
vif(model)
```

```
## pctwhite pctblack pctasian
## 4.501114 4.193772 1.291071
```

```
summary(df$pctotherrace)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.2867   0.7826   2.0031   2.1066  41.9303
```

```
hist(df$pctotherrace, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctotherrace), sd(df$pctotherrace)), add = T)
```




```
shapiro.test(df$pctotherrace)
```

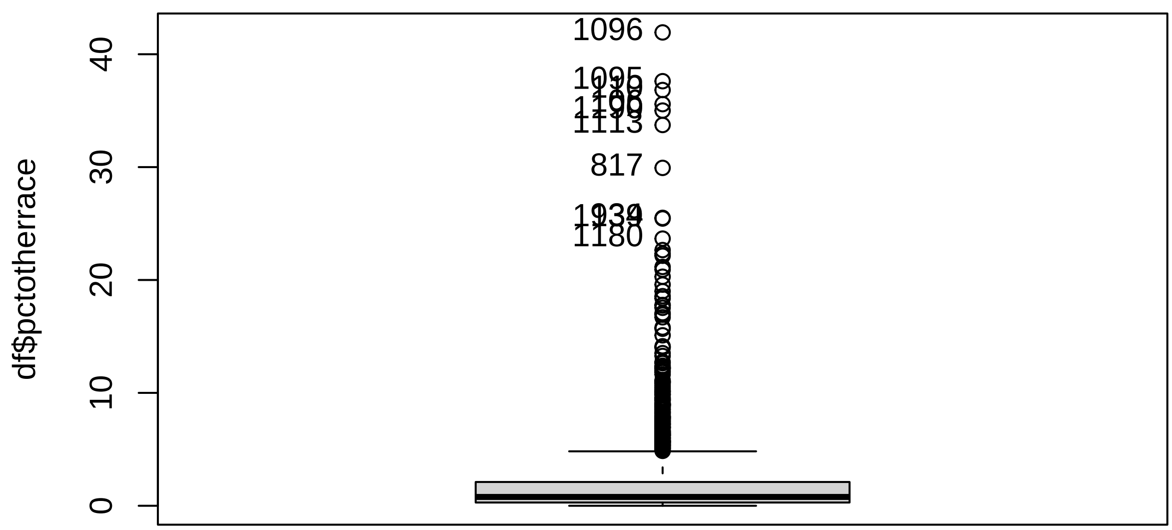
```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  df$pctotherrace
## W = 0.50981, p-value < 2.2e-16
```

```r
sum(is.na(df$pctotherrace))
```

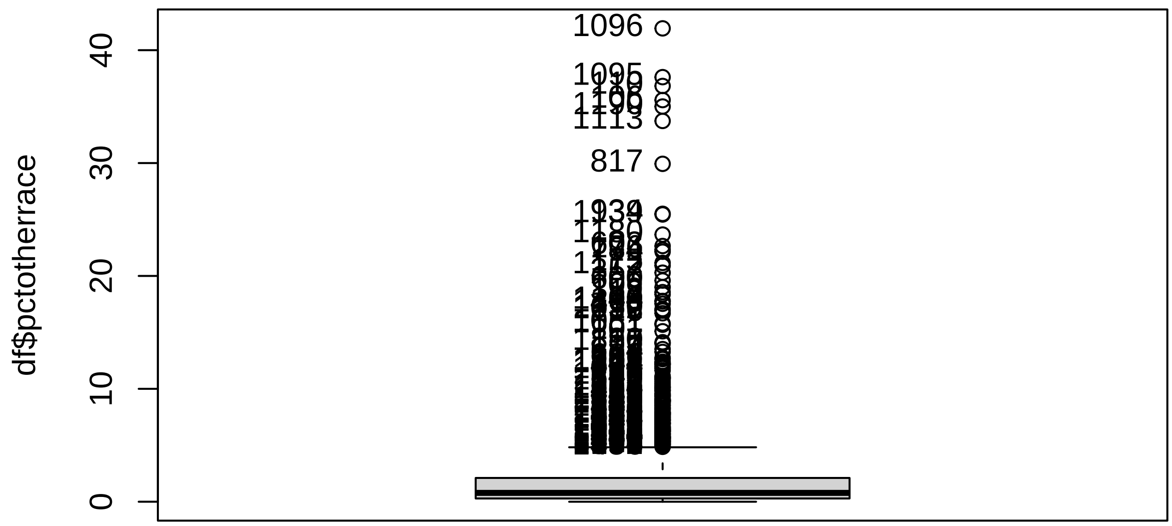
```
## [1] 0
```

```r
Boxplot(df$pctotherrace)
```




```
##  [1] 1096 1095  110  106 1190 1113  817  934 1139 1180
```

```r
length(Boxplot(df$pctotherrace, id = list(n=Inf)))
```



```
## [1] 181
```

```
sevout_pctotherrace = (quantile(df$pctotherrace,0.25)+(3*((quantile(df$pctotherrace,0.75)-quantile(df$pc
length(which(df$pctotherrace > sevout_pctotherrace))
```

```
## [1] 148
```

```
df$f.pctotherrace <- ifelse(df$pctotherrace <= 0.2867, 1, ifelse(df$pctotherrace > 0.2867 & df$pctother
df$f.pctotherrace <- factor(df$f.pctotherrace, labels=c("LowOtherRace%","LowMidOtherRace%","HighMidOthe
table(df$f.pctotherrace)
```

```
##
##     LowOtherRace%  LowMidOtherRace% HighMidOtherRace%    HighOtherRace%
##               458               458               457               458
```

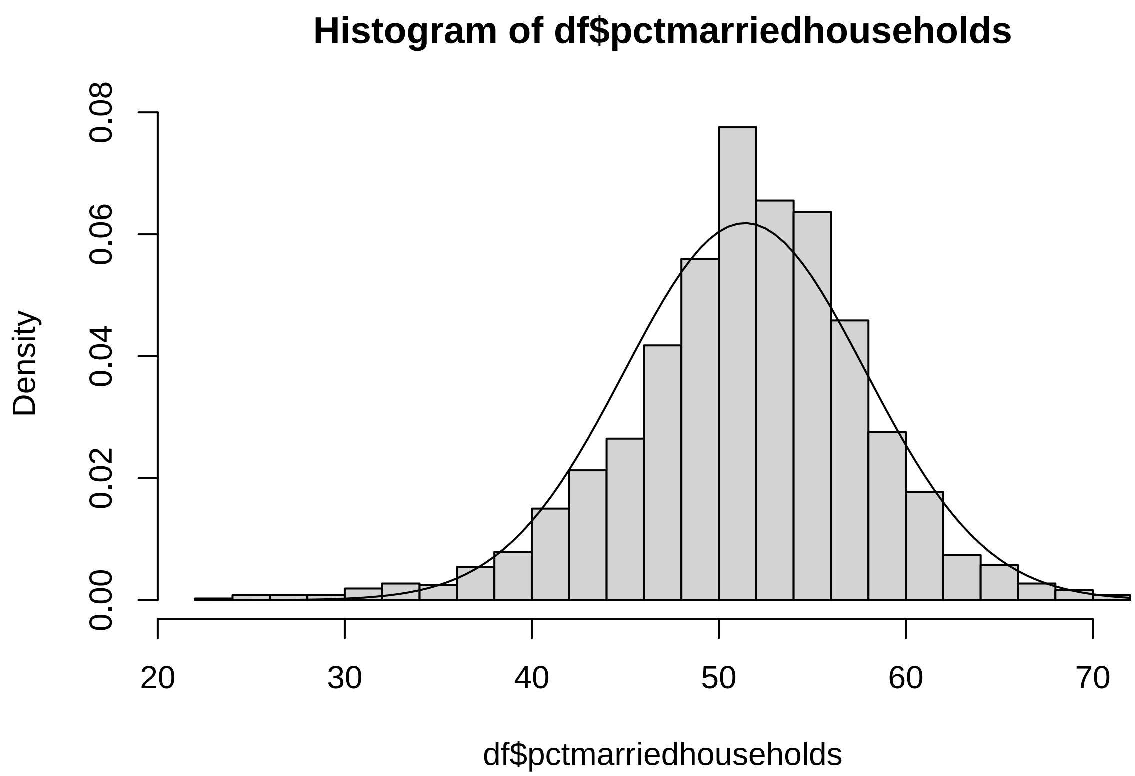
Variable 29 - pctmarriedhouseholds

Another continuous ratio variable not normally distributed with 0 missing values, 57 outliers (2 of them severe) on both ends of the spectrum. We create an additional ordinal factor "f.pctmarriedhouseholds".

```
summary(df$pctmarriedhouseholds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.99   47.85   51.73   51.40   55.48   71.40
```

```r
hist(df$pctmarriedhouseholds, breaks = 30, freq = F)
curve(dnorm(x, mean(df$pctmarriedhouseholds), sd(df$pctmarriedhouseholds)), add = T)
```



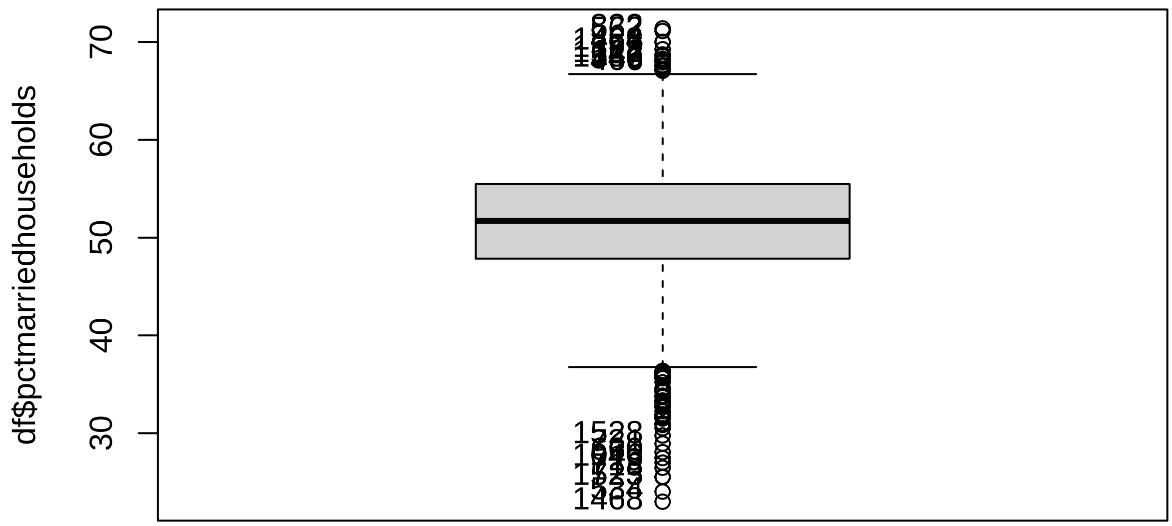

```r
shapiro.test(df$pctmarriedhouseholds)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$pctmarriedhouseholds
## W = 0.9816, p-value = 1.341e-14
```

```r
sum(is.na(df$pctmarriedhouseholds))
```

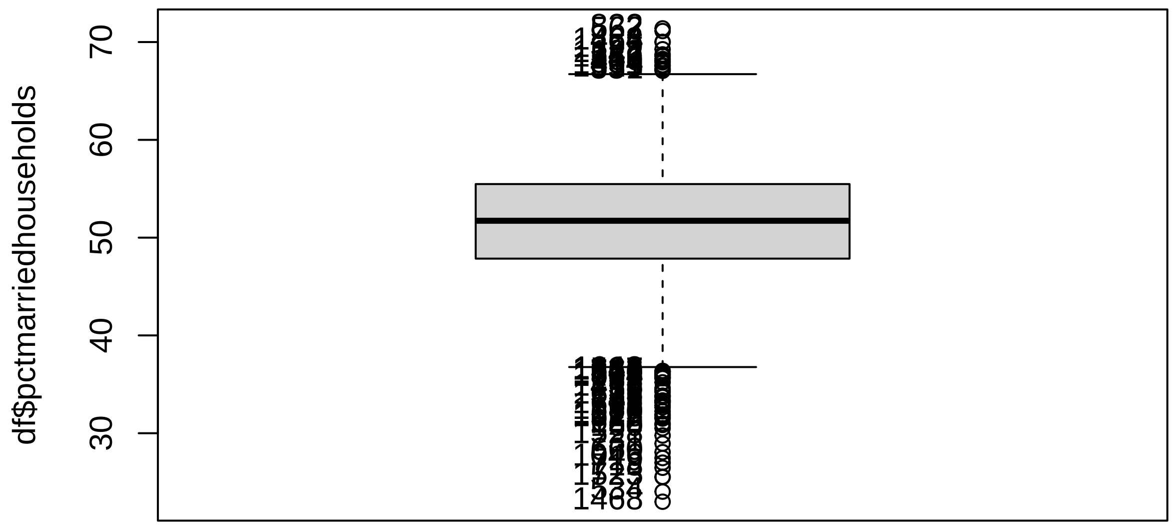
```
## [1] 0
```

```r
Boxplot(df$pctmarriedhouseholds)
```




```
##  [1] 1468  534 1525  723  718  719 1046  660  731 1528  822  562 1423  464 1399
## [16] 1122  547 1556  549  466
```

```r
length(Boxplot(df$pctmarriedhouseholds, id = list(n=Inf)))
```



```
## [1] 57
```

```
sevout_pctmarriedhouseholds = (quantile(df$pctmarriedhouseholds,0.25)+(3*((quantile(df$pctmarriedhouseho
length(which(df$pctmarriedhouseholds > sevout_pctmarriedhouseholds))
```

```
## [1] 2
```

```
df$f.pctmarriedhouseholds <- ifelse(df$pctmarriedhouseholds <= 47.85, 1, ifelse(df$pctmarriedhouseholds
df$f.pctmarriedhouseholds <- factor(df$f.pctmarriedhouseholds, labels=c("LowMarried%","LowMidMarried%",
table(df$f.pctmarriedhouseholds)
```

```
##
##     LowMarried%  LowMidMarried% HighMidMarried%    HighMarried%
##             457             460             456             458
```

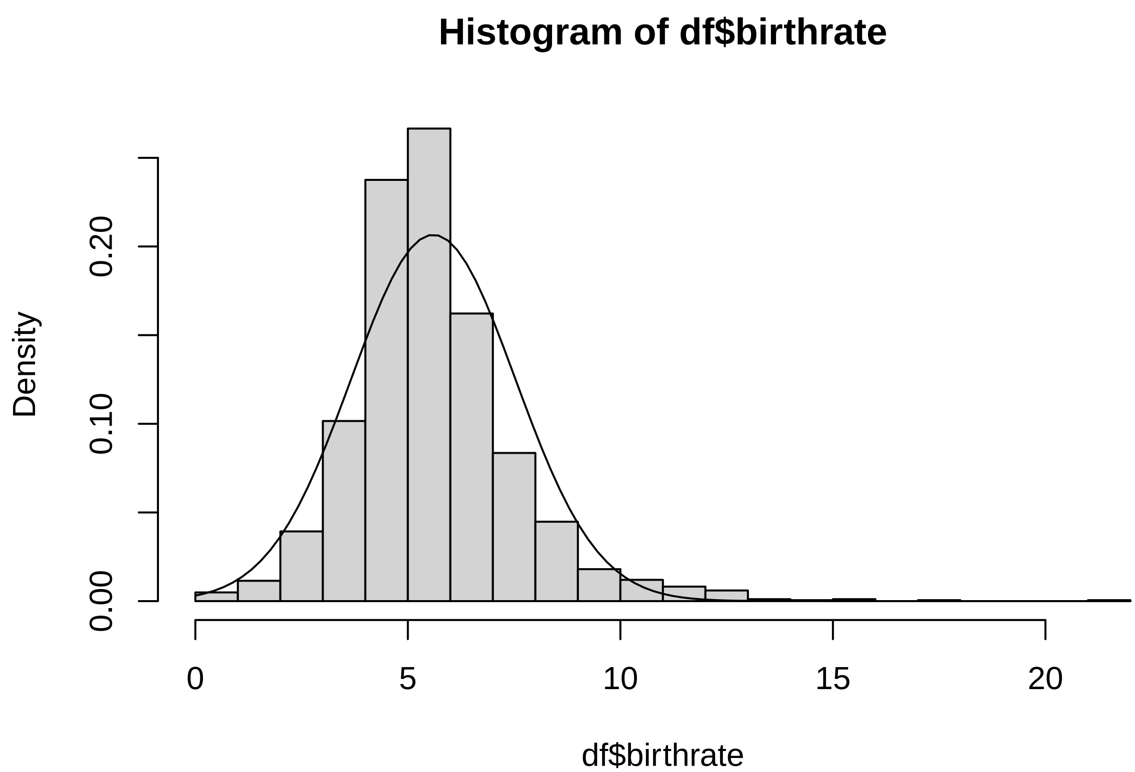
Variable 30 - birthrate

The last variable is yet another continuous ratio variable not normally distributed with 0 missing values, 104 outliers (52 of them severe) on both ends of the spectrum. We create an additional ordinal factor "f.birthrate".

```
summary(df$birthrate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   4.528   5.355   5.597   6.414  21.326
```

```r
hist(df$birthrate, breaks = 30, freq = F)
curve(dnorm(x, mean(df$birthrate), sd(df$birthrate)), add = T)
```



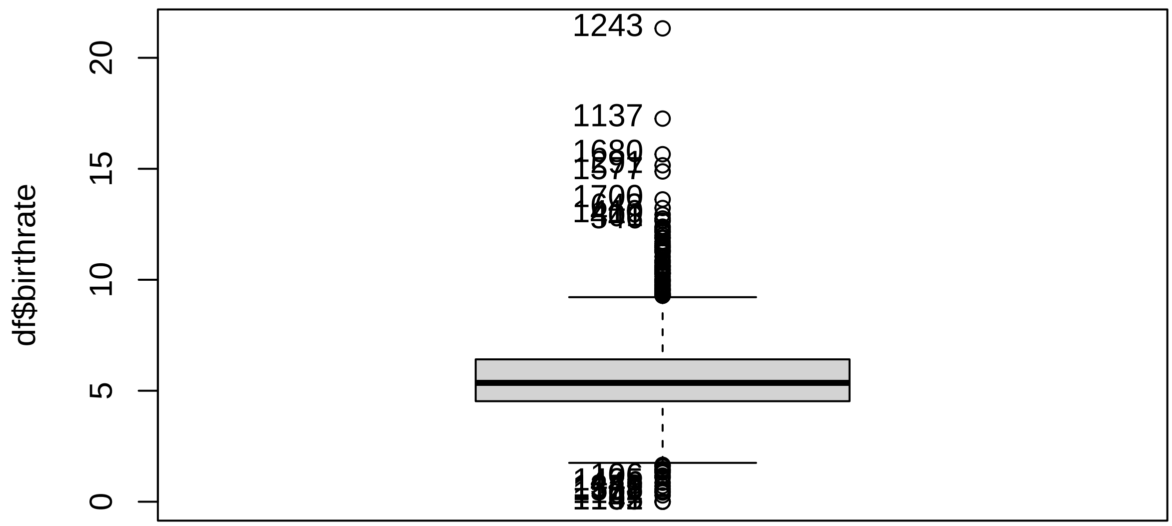

```r
shapiro.test(df$birthrate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$birthrate
## W = 0.93107, p-value < 2.2e-16
```

```r
sum(is.na(df$birthrate))
```

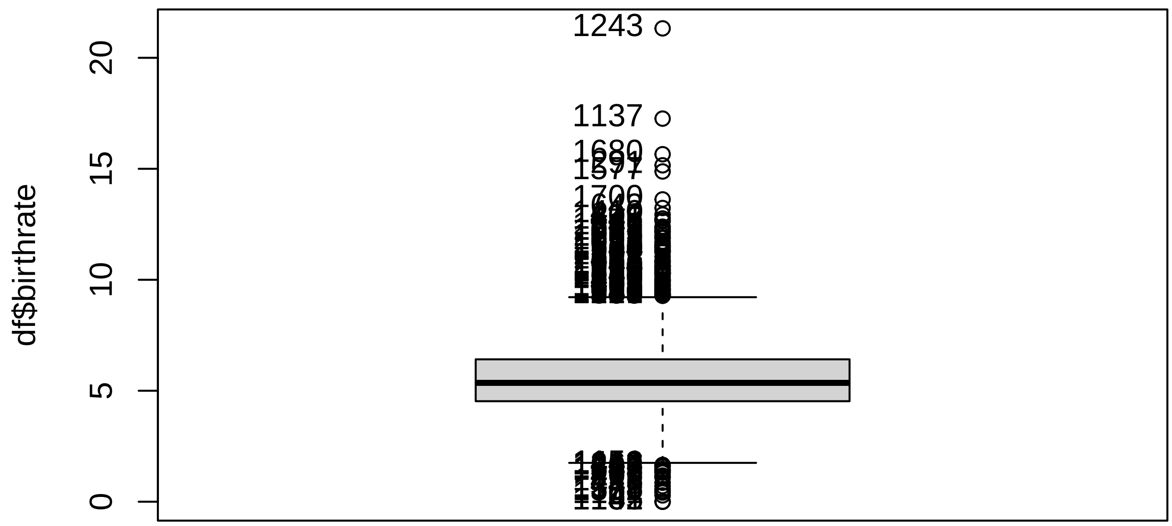
```
## [1] 0
```

```r
Boxplot(df$birthrate)
```




```
## [1]  101 1135 1142 1101 1522 1378 1373  446 1425  106 1243 1137 1680  291 1577
## [16] 1700  643 1410  401  546
```

```r
length(Boxplot(df$birthrate, id = list(n=Inf)))
```




```
## [1] 104
```

```r
sevout_birthrate = (quantile(df$birthrate,0.25)+(3*((quantile(df$birthrate,0.75)-quantile(df$birthrate,
length(which(df$birthrate > sevout_birthrate))
```

```
## [1] 52
```

```r
df$f.birthrate <- ifelse(df$birthrate <= 4.528, 1, ifelse(df$birthrate > 4.528 & df$birthrate <= 5.355,
df$f.birthrate <- factor(df$f.birthrate, labels=c("LowBirth%","LowMidBirth%","HighMidBirth%","HighBirth
table(df$f.birthrate)
```

```
##
##     LowBirth%  LowMidBirth% HighMidBirth%    HighBirth%
##           458           458           456           459
```

Missing data

There is only one variable left with missing data, pctemployed16_over with 82 NAs. Since the number is low and a priori this variable can be useful so we will fix missing data using the mice method. We will also update "f.pctemployed16_over" with the new imputed data but the same quartile limits as before.

```r
res.mice <- mice(df)
```

```
##
##  iter imp variable
##   1   1  pctemployed16_over  f.pctemployed16_over
##   1   2  pctemployed16_over  f.pctemployed16_over
##   1   3  pctemployed16_over  f.pctemployed16_over
##   1   4  pctemployed16_over  f.pctemployed16_over
##   1   5  pctemployed16_over  f.pctemployed16_over
##   2   1  pctemployed16_over  f.pctemployed16_over
##   2   2  pctemployed16_over  f.pctemployed16_over
##   2   3  pctemployed16_over  f.pctemployed16_over
##   2   4  pctemployed16_over  f.pctemployed16_over
##   2   5  pctemployed16_over  f.pctemployed16_over
##   3   1  pctemployed16_over  f.pctemployed16_over
##   3   2  pctemployed16_over  f.pctemployed16_over
##   3   3  pctemployed16_over  f.pctemployed16_over
##   3   4  pctemployed16_over  f.pctemployed16_over
##   3   5  pctemployed16_over  f.pctemployed16_over
##   4   1  pctemployed16_over  f.pctemployed16_over
##   4   2  pctemployed16_over  f.pctemployed16_over
##   4   3  pctemployed16_over  f.pctemployed16_over
##   4   4  pctemployed16_over  f.pctemployed16_over
##   4   5  pctemployed16_over  f.pctemployed16_over
##   5   1  pctemployed16_over  f.pctemployed16_over
##   5   2  pctemployed16_over  f.pctemployed16_over
##   5   3  pctemployed16_over  f.pctemployed16_over
##   5   4  pctemployed16_over  f.pctemployed16_over
##   5   5  pctemployed16_over  f.pctemployed16_over
```

```
## Warning: Number of logged events: 3
```

```r
df$pctemployed16_over <- complete(res.mice, action = 1)$pctemployed16_over

df$f.pctemployed16_over <- ifelse(df$pctemployed16_over <= 48.60, 1, ifelse(df$pctemployed16_over > 48.
df$f.pctemployed16_over <- factor(df$f.pctemployed16_over, labels=c("LowEmploy%","LowMidEmploy%","HighM
table(df$f.pctemployed16_over)
```

```
##
##     LowEmploy%  LowMidEmploy% HighMidEmploy%    HighEmploy%
##            460            459            467            445
```

Duplicate Removal

Since we have a variable with unique values for each row (geography), we can check for duplicates easily by counting unique values for geography and comparing with the number of observations of our data. Since there is no difference there are no duplicates.

```r
nrow(df)
```

```
## [1] 1831
```

```r
length(unique(df$geography))
```

```
## [1] 1831
```

Outliers

For each observation we will count how many times it is an outlier of a numerical variable. We will add the count to a new variable called "univariate_outlier_count". If we look at the individuals that are outliers in 10 or more variables we have a total of 8 counties. All of them have high percentages of non-white population, both black and asian, a low median age, a high mortality count and a high bias towards private and employee health coverage. Of these 8 counties, 6 are wealthy (Low poverty percent) and 2 are poor. It is chosen to delete these outliers from the data set for the rest of the project.

```r
count_outliers <- function(data) {
  # Function to check for outliers based on IQR
  is_outlier <- function(x) {
    Q1 <- quantile(x, 0.25, na.rm = TRUE)
    Q3 <- quantile(x, 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1
    lower_bound <- Q1 - 1.5 * IQR
    upper_bound <- Q3 + 1.5 * IQR
    return(x < lower_bound | x > upper_bound)
  }

  # Apply the outlier function to each column and sum the results for each row using dplyr
  data %>%
    mutate(outlier_count = rowSums(sapply(., is_outlier), na.rm = TRUE))
}

df$univariate_outlier_count <- count_outliers(df[, c(1:12, 14:31)])$outlier_count
table(df$univariate_outlier_count)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  12
## 742 479 217 122  85  66  50  30  21  11   7   1
```

```r
df[which(df$univariate_outlier_count >= 10),]
```

```
##      avganncount avgdeathsperyear target_deathrate incidencerate medincome
## 128      862.000              283            136.5      364.9000    122641
## 168      135.000               23            162.1     1014.2000     46954
## 529     4139.000             1292            120.1      392.9000     97279
## 613     3648.000             1186            140.0      447.0000    100806
## 615     7334.000             2355            135.0      420.0000     97219
## 792     1962.668              796            146.8      453.5494     76104
## 1046    8236.000             3303            211.7      533.5000     39037
## 1309     954.000              327            146.5      397.9000     89861
##      popest2015 povertypercent studypercap binnedinc medianage medianagemale
## 128      375629            3.9    449.9120  93564.75      35.3          34.9
## 168       15052           20.1      0.0000  46611.30      24.6          25.6
## 529     1040116            7.2    184.5948  93564.75      38.5          37.0
```

```
## 613      765135           7.5    218.2621  93564.75      39.5          38.2
## 615     1918044           8.5    410.3138  93564.75      36.8          35.9
## 792      580159           6.6    449.8767  93564.75      36.8          35.6
## 1046    1567442          25.8    742.6112  38888.25      33.7          32.2
## 1309     309697           4.9    129.1585  93564.75      36.1          35.5
##      medianagefemale                          geography percentmarried
## 128             35.6         Loudoun County, Virginia           61.2
## 168             23.6       Williamsburg city, Virginia           26.2
## 529             40.0       Montgomery County, Maryland           53.2
## 613             40.8      San Mateo County, California           51.9
## 615             37.8    Santa Clara County, California           53.2
## 792             38.1            Johnson County, Kansas           56.8
## 1046            35.2 Philadelphia County, Pennsylvania           29.3
## 1309            36.7          Hamilton County, Indiana           62.3
##      pctnohs18_24 pcths18_24 pctbachdeg18_24 pcths25_over pctbachdeg25_over
## 128          16.6       26.5            17.1         13.8              34.8
## 168           1.5       10.0            10.2         15.5              27.1
## 529          12.7       23.5            19.9         14.0              26.6
## 613          11.7       25.5            16.2         16.5              27.1
## 615          10.6       25.8            16.8         15.2              26.1
## 792          11.5       25.0            17.1         15.2              33.6
## 1046         14.3       30.1            12.6         33.8              14.9
## 1309         18.4       27.1            19.7         15.9              35.5
##      pctemployed16_over pctunemployed16_over pctprivatecoverage
## 128                72.6                  4.0               86.9
## 168                44.5                  8.5               83.3
## 529                67.1                  6.1               77.0
## 613                64.1                  6.7               76.0
## 615                61.9                  7.7               74.1
## 792                69.2                  4.5               84.0
## 1046               51.4                 13.9               55.7
## 1309               70.1                  4.3               86.4
##      pctempprivcoverage pctpubliccoverage pctpubliccoveragealone pctwhite
## 128                68.9              11.8                    4.6 67.77025
## 168                52.2              22.0                    8.9 74.88817
## 529                56.4              23.0                   11.5 55.62676
## 613                55.7              25.8                   13.2 54.97635
## 615                57.3              24.8                   14.2 48.30471
## 792                63.0              18.9                    8.0 86.91211
## 1046               38.8              41.3                   27.6 41.67215
## 1309               68.8              14.8                    6.2 87.62182
##       pctblack  pctasian pctotherrace pctmarriedhouseholds birthrate
## 128   7.432026 16.200029    3.6257330             65.51326  6.198748
## 168  15.277213  5.889928    0.4608920             36.33759  2.181467
## 529  17.607940 14.561938    7.8599295             53.70241  5.281995
## 613   2.596260 26.558136    9.4474518             53.65425  5.015576
## 615   2.585982 33.760905    9.8342798             56.30311  5.541785
## 792   4.488774  4.460193    0.9182907             55.46135  5.529393
## 1046 42.757570  6.864827    5.5732468             27.45994  5.282606
## 1309  3.568358  5.348661    0.9071755             62.29758  5.756462
##        f.avganncount f.avgdeathsperyear     f.deathrate    f.incidencerate
## 128    HighCaseCount      HighMortCount    LowDeathrate     LowDiagnPerCap
## 168  LowMidCaseCount       LowMortCount LowMidDeathrate    HighDiagnPerCap
## 529    HighCaseCount      HighMortCount    LowDeathrate     LowDiagnPerCap
```

```
## 613    HighCaseCount      HighMortCount    LowDeathrate  LowMidDiagnPerCap
## 615    HighCaseCount      HighMortCount    LowDeathrate     LowDiagnPerCap
## 792    HighCaseCount      HighMortCount    LowDeathrate HighMidDiagnPerCap
## 1046   HighCaseCount      HighMortCount   HighDeathrate    HighDiagnPerCap
## 1309   HighCaseCount      HighMortCount    LowDeathrate     LowDiagnPerCap
##           f.medincome f.popest2015 f.povertypercent f.studypercap
## 128     HighMedianInc      HighPop          LowPov%    HighTrials
## 168  HighMidMedianInc    LowMidPop      HighMidPov%      NoTrials
## 529     HighMedianInc      HighPop          LowPov%    HighTrials
## 613     HighMedianInc      HighPop          LowPov%    HighTrials
## 615     HighMedianInc      HighPop          LowPov%    HighTrials
## 792     HighMedianInc      HighPop          LowPov%    HighTrials
## 1046  LowMidMedianInc      HighPop         HighPov%    HighTrials
## 1309    HighMedianInc      HighPop          LowPov%     MidTrials
##           f.binnedinc f.medianage f.medianagemale f.medianagefemale
## 128     HighIncPerCap      LowAge      LowAgeMale      LowAgeFemale
## 168  HighMidIncPerCap      LowAge      LowAgeMale      LowAgeFemale
## 529     HighIncPerCap   LowMidAge   LowMidAgeMale   LowMidAgeFemale
## 613     HighIncPerCap   LowMidAge   LowMidAgeMale   LowMidAgeFemale
## 615     HighIncPerCap      LowAge      LowAgeMale      LowAgeFemale
## 792     HighIncPerCap      LowAge      LowAgeMale      LowAgeFemale
## 1046  LowMidIncPerCap      LowAge      LowAgeMale      LowAgeFemale
## 1309    HighIncPerCap      LowAge      LowAgeMale      LowAgeFemale
##             state f.percentmarried   f.pctnohs18_24  f.pcths18_24
## 128      Virginia    HighMarriage%  LowMidNoHighsc%    LowHighsc%
## 168      Virginia     LowMarriage%     LowNoHighsc%    LowHighsc%
## 529      Maryland HighMidMarriage%     LowNoHighsc%    LowHighsc%
## 613    California  LowMidMarriage%     LowNoHighsc%    LowHighsc%
## 615    California HighMidMarriage%     LowNoHighsc%    LowHighsc%
## 792        Kansas    HighMarriage%     LowNoHighsc%    LowHighsc%
## 1046 Pennsylvania     LowMarriage%  LowMidNoHighsc% LowMidHighsc%
## 1309      Indiana    HighMarriage% HighMidNoHighsc%    LowHighsc%
##       f.pcths25_over f.pctbachdeg25_over f.pctemployed16_over
## 128     Low25Highsc%           HighBach%          HighEmploy%
## 168     Low25Highsc%           HighBach%           LowEmploy%
## 529     Low25Highsc%           HighBach%          HighEmploy%
## 613     Low25Highsc%           HighBach%          HighEmploy%
## 615     Low25Highsc%           HighBach%          HighEmploy%
## 792     Low25Highsc%           HighBach%          HighEmploy%
## 1046 LowMid25Highsc%        HighMidBach%        LowMidEmploy%
## 1309    Low25Highsc%           HighBach%          HighEmploy%
##      f.pctunemployed16_over f.pcuntemployed16_over f.pctprivatecoverage
## 128                       1            LowUnEmploy%         HighPrivate%
## 168                       3        HighMidUnEmploy%         HighPrivate%
## 529                       2         LowMidUnEmploy%         HighPrivate%
## 613                       2         LowMidUnEmploy%         HighPrivate%
## 615                       3        HighMidUnEmploy%         HighPrivate%
## 792                       1            LowUnEmploy%         HighPrivate%
## 1046                      4           HighUnEmploy%          LowPrivate%
## 1309                      1            LowUnEmploy%         HighPrivate%
##       f.pctempprivcoverage f.pctpubliccoverage f.pctpubliccoveragealone
## 128    HighEmployeeHealth%       LowGovHealth%       LowGovHealthAlone%
## 168    HighEmployeeHealth%       LowGovHealth%       LowGovHealthAlone%
## 529    HighEmployeeHealth%       LowGovHealth%       LowGovHealthAlone%
```

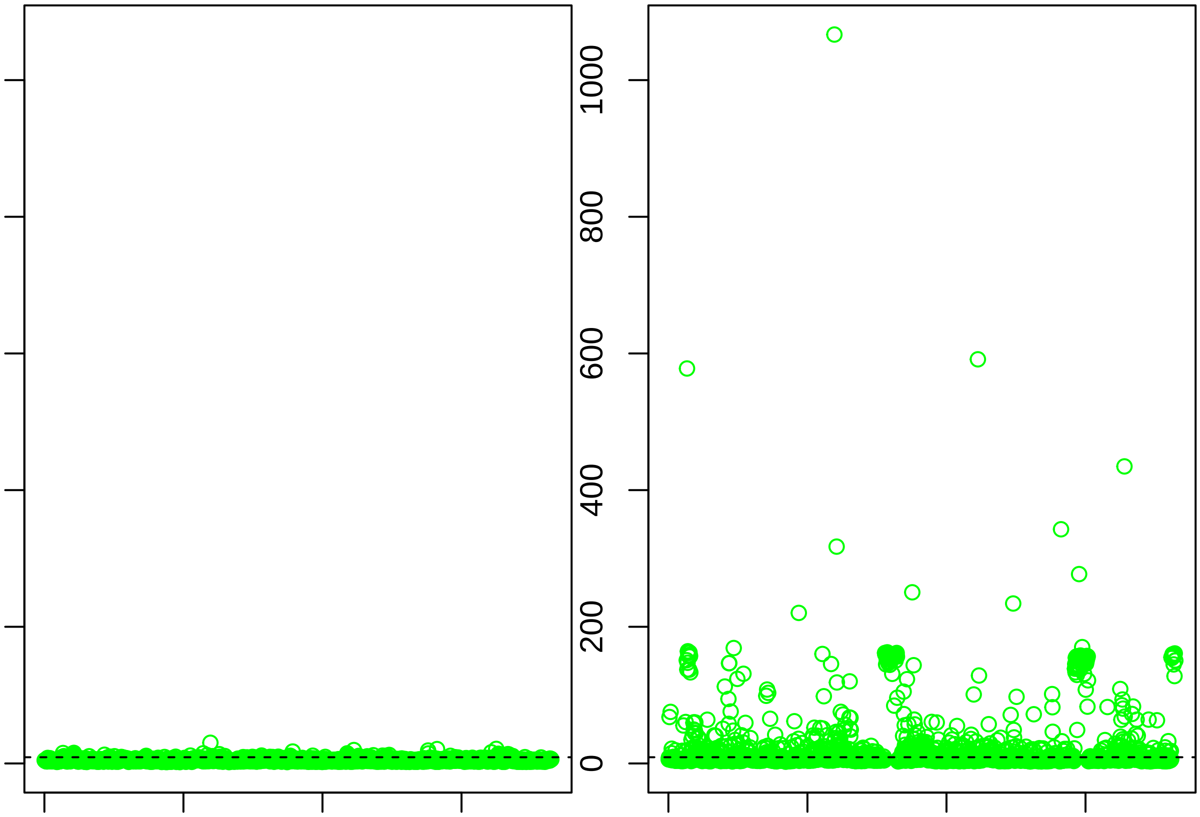
```
## 613    HighEmployeeHealth%       LowGovHealth%       LowGovHealthAlone%
## 615    HighEmployeeHealth%       LowGovHealth%       LowGovHealthAlone%
## 792    HighEmployeeHealth%       LowGovHealth%       LowGovHealthAlone%
## 1046 LowMidEmployeeHealth%   HighMidGovHealth%      HighGovHealthAlone%
## 1309   HighEmployeeHealth%       LowGovHealth%       LowGovHealthAlone%
##        f.pctwhite    f.pctblack f.pctasian    f.pctotherrace
## 128     LowWhite% HighMidBlack% HighAsian%    HighOtherRace%
## 168     LowWhite%    HighBlack% HighAsian%  LowMidOtherRace%
## 529     LowWhite%    HighBlack% HighAsian%    HighOtherRace%
## 613     LowWhite% HighMidBlack% HighAsian%    HighOtherRace%
## 615     LowWhite% HighMidBlack% HighAsian%    HighOtherRace%
## 792  LowMidWhite% HighMidBlack% HighAsian% HighMidOtherRace%
## 1046    LowWhite%    HighBlack% HighAsian%    HighOtherRace%
## 1309 LowMidWhite% HighMidBlack% HighAsian% HighMidOtherRace%
##      f.pctmarriedhouseholds   f.birthrate univariate_outlier_count
## 128            HighMarried% HighMidBirth%                       10
## 168             LowMarried%     LowBirth%                       10
## 529         HighMidMarried%  LowMidBirth%                       10
## 613         HighMidMarried%  LowMidBirth%                       10
## 615            HighMarried% HighMidBirth%                       12
## 792         HighMidMarried% HighMidBirth%                       10
## 1046            LowMarried%  LowMidBirth%                       10
## 1309           HighMarried% HighMidBirth%                       10
```

```
df = df[-which(df$univariate_outlier_count >= 10),]
```

Multivariate Outliers

We will apply Moutlier on the numerical variables in order to find multivariate outliers. We have to perform the calculation excluding the variable studypercap because otherwise the method is unable to execute due to multicollinearity casuing a singularity matrix in the intermediate calculations. An extremely mild threshold is chosen (0.00005%) because even using this threshold we get a significant amount of multivariate outliers, 4% of the total sample. Lowering the threshold even further doesn't change much the amount of outliers and rising it higher makes the amount of outliers rise too much (10% outliers at 0.1% significance level). We also choose to delete these outliers from the data set for the rest of the project.

```
par(mar = c(1, 1, 1, 1))
res.out = Moutlier(df[, c(1:7,9:12,14:31)], quantile = 0.9999995, col="green")
```




```r
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))
```

```
##   62   67   70   71   76   94  101  103  104  106  110  162  218  237  254  279
##   62   67   70   71   76   94  101  103  104  106  110  161  216  235  252  277
##  364  368  434  474  527  574  588  600  608  634  636  654  720  753  786  817
##  362  366  432  472  525  571  585  597  605  629  631  649  715  748  781  811
##  827  847  883  890  892  899  971 1016 1091 1094 1095 1096 1113 1120 1137 1139
##  821  841  877  884  886  893  965 1010 1084 1087 1088 1089 1106 1113 1130 1132
## 1168 1180 1190 1223 1238 1243 1247 1388 1389 1420 1468 1485 1615 1633 1636 1637
## 1161 1173 1183 1216 1231 1236 1240 1380 1381 1412 1460 1477 1607 1625 1628 1629
## 1639 1641 1643 1648 1675 1679 1680 1681 1692 1736
## 1631 1633 1635 1640 1667 1671 1672 1673 1684 1728
```

```r
length(which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))
)/1823
```

```
## [1] 0.04059243
```

```r
plot( res.out$md, res.out$rd )
abline(h=res.out$cutoff, col="red")
abline(v=res.out$cutoff, col="red")

summary(df[which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff)),])
```

```
##   avganncount       avgdeathsperyear  target_deathrate incidencerate
##  Min.   :    7.00   Min.   :    3.0   Min.   : 59.7    Min.   : 201.3
##  1st Qu.:   20.25   1st Qu.:    7.0   1st Qu.:144.9    1st Qu.: 382.6
##  Median :   58.50   Median :   15.0   Median :168.8    Median : 427.5
##  Mean   : 2650.56   Mean   :  952.8   Mean   :175.4    Mean   : 434.8
##  3rd Qu.: 1962.67   3rd Qu.:  241.2   3rd Qu.:198.4    3rd Qu.: 466.1
##  Max.   :38150.00   Max.   :14010.0   Max.   :362.8    Max.   :1206.9
##    medincome        popest2015       povertypercent   studypercap
##  Min.   : 27627   Min.   :     829   Min.   : 3.70   Min.   :   0.0
##  1st Qu.: 37200   1st Qu.:    3974   1st Qu.:12.45   1st Qu.:   0.0
##  Median : 46897   Median :    8862   Median :17.20   Median :   0.0
##  Mean   : 52029   Mean   :  616179   Mean   :18.63   Mean   : 178.9
##  3rd Qu.: 59047   3rd Qu.:  218451   3rd Qu.:24.02   3rd Qu.: 165.7
##  Max.   :110507   Max.   :10170292   Max.   :41.90   Max.   :3046.5
##    binnedinc       medianage     medianagemale   medianagefemale
##  Min.   :28429   Min.   :23.30   Min.   :23.00   Min.   :24.50
##  1st Qu.:36584   1st Qu.:34.17   1st Qu.:33.27   1st Qu.:34.05
##  Median :46611   Median :38.05   Median :36.65   Median :40.10
##  Mean   :53564   Mean   :38.20   Mean   :37.16   Mean   :39.61
##  3rd Qu.:58020   3rd Qu.:41.60   3rd Qu.:41.42   3rd Qu.:44.05
##  Max.   :93565   Max.   :56.50   Max.   :58.60   Max.   :55.00
##   geography         percentmarried   pctnohs18_24     pcths18_24
##  Length:74          Min.   :23.10   Min.   : 0.50   Min.   : 0.00
##  Class :character   1st Qu.:40.62   1st Qu.:12.03   1st Qu.:25.30
##  Mode  :character   Median :45.50   Median :18.25   Median :31.30
##                     Mean   :47.07   Mean   :22.11   Mean   :33.44
##                     3rd Qu.:54.65   3rd Qu.:30.35   3rd Qu.:42.17
##                     Max.   :66.60   Max.   :59.10   Max.   :72.50
##  pctbachdeg18_24   pcths25_over   pctbachdeg25_over pctemployed16_over
##  Min.   : 0.000   Min.   : 8.30   Min.   : 4.400    Min.   :24.00
##  1st Qu.: 1.175   1st Qu.:24.98   1st Qu.: 9.225    1st Qu.:45.62
##  Median : 4.550   Median :30.35   Median :13.450    Median :54.45
##  Mean   : 8.243   Mean   :29.94   Mean   :15.505    Mean   :52.99
##  3rd Qu.:10.875   3rd Qu.:36.02   3rd Qu.:19.200    3rd Qu.:62.02
##  Max.   :51.800   Max.   :44.60   Max.   :42.200    Max.   :80.10
##  pctunemployed16_over pctprivatecoverage pctempprivcoverage pctpubliccoverage
##  Min.   : 0.700       Min.   :27.80      Min.   :14.30      Min.   :11.20
##  1st Qu.: 4.825       1st Qu.:47.60      1st Qu.:27.10      1st Qu.:26.20
##  Median : 7.500       Median :59.85      Median :36.95      Median :35.50
##  Mean   : 8.899       Mean   :59.27      Mean   :37.47      Mean   :34.26
##  3rd Qu.:11.600       3rd Qu.:72.67      3rd Qu.:45.77      3rd Qu.:42.40
##  Max.   :29.400       Max.   :89.60      Max.   :69.20      Max.   :62.70
##  pctpubliccoveragealone    pctwhite        pctblack          pctasian
##  Min.   : 2.60          Min.   :12.27   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.:12.75          1st Qu.:53.51   1st Qu.: 0.1552   1st Qu.: 0.1579
##  Median :19.95          Median :68.49   Median : 1.5478   Median : 1.2747
##  Mean   :19.99          Mean   :65.87   Mean   : 8.1935   Mean   : 4.7156
##  3rd Qu.:26.30          3rd Qu.:84.72   3rd Qu.:11.2534   3rd Qu.: 4.0873
##  Max.   :46.60          Max.   :98.47   Max.   :65.1433   Max.   :37.1569
##   pctotherrace     pctmarriedhouseholds   birthrate               f.avganncount
##  Min.   : 0.0000   Min.   :22.99        Min.   : 0.000   LowCaseCount    :40
##  1st Qu.: 0.6687   1st Qu.:41.65        1st Qu.: 4.673   LowMidCaseCount : 4
##  Median : 2.7677   Median :46.98        Median : 5.343   HighMidCaseCount: 3
##  Mean   : 7.4390   Mean   :48.05        Mean   : 6.378   HighCaseCount   :27
```

```
##  3rd Qu.: 8.8884   3rd Qu.:55.64        3rd Qu.: 6.862
##  Max.   :41.9303   Max.   :67.26        Max.   :21.326
##        f.avgdeathsperyear           f.deathrate            f.incidencerate
##  LowMortCount    :45       LowDeathrate    :30   LowDiagnPerCap    :33
##  LowMidMortCount : 6       LowMidDeathrate :14   LowMidDiagnPerCap :10
##  HighMidMortCount: 2       HighMidDeathrate: 7   HighMidDiagnPerCap:17
##  HighMortCount   :21       HighDeathrate   :23   HighDiagnPerCap   :14
##
##
##           f.medincome     f.popest2015    f.povertypercent    f.studypercap
##  LowMedianInc    :21   LowPop    :40    LowPov%    :18      NoTrials  :48
##  LowMidMedianInc :14   LowMidPop :10    LowMidPov% :16      MidTrials : 6
##  HighMidMedianInc:11   HighMidPop: 2    HighMidPov%:12      HighTrials:20
##  HighMedianInc   :28   HighPop   :22    HighPov%   :28
##
##
##           f.binnedinc     f.medianage       f.medianagemale
##  LowIncPerCap    :19   LowAge    :36   LowAgeMale    :37
##  LowMidIncPerCap :16   LowMidAge :16   LowMidAgeMale :12
##  HighMidIncPerCap:15   HighMidAge: 9   HighMidAgeMale:12
##  HighIncPerCap   :24   HighAge   :13   HighAgeMale   :13
##
##
##        f.medianagefemale    state                  f.percentmarried
##  LowAgeFemale    :35      Length:74          LowMarriage%    :44
##  LowMidAgeFemale :11      Class :character   LowMidMarriage% : 8
##  HighMidAgeFemale:14      Mode  :character   HighMidMarriage%: 8
##  HighAgeFemale   :14                         HighMarriage%   :14
##
##
##          f.pctnohs18_24         f.pcths18_24          f.pcths25_over
##  LowNoHighsc%    :23     LowHighsc%    :31    Low25Highsc%    :37
##  LowMidNoHighsc% :13     LowMidHighsc% :12    LowMid25Highsc% :13
##  HighMidNoHighsc%: 6     HighMidHighsc%:10    HighMid25Highsc%:16
##  HighNoHighsc%   :32     HighHighsc%   :21    High25Highsc%   : 8
##
##
##    f.pctbachdeg25_over     f.pctemployed16_over f.pctunemployed16_over
##  LowBach%    :20       LowEmploy%    :23        Min.   :1.000
##  LowMidBach% :12       LowMidEmploy% :14        1st Qu.:1.000
##  HighMidBach%:12       HighMidEmploy%:12        Median :2.000
##  HighBach%   :30       HighEmploy%   :25        Mean   :2.514
##                                                 3rd Qu.:4.000
##                                                 Max.   :4.000
##      f.pcuntemployed16_over      f.pctprivatecoverage
##  LowUnEmploy%    :25         LowPrivate%    :28
##  LowMidUnEmploy% :13         LowMidPrivate% :22
##  HighMidUnEmploy%: 9         HighMidPrivate%: 3
##  HighUnEmploy%   :27         HighPrivate%   :21
##
##
##             f.pctempprivcoverage        f.pctpubliccoverage
##  LowEmployeeHealth%    :34        LowGovHealth%    :28
##  LowMidEmployeeHealth% :14        LowMidGovHealth% :13
```

```
##  HighMidEmployeeHealth%:11        HighMidGovHealth%:12
##  HighEmployeeHealth%   :15        HighGovHealth%   :21
##
##
##           f.pctpubliccoveragealone         f.pctwhite         f.pctblack
##  LowGovHealthAlone%    :24          LowWhite%    :49   LowBlack%    :29
##  LowMidGovHealthAlone% : 9          LowMidWhite% :12   LowMidBlack% :11
##  HighMidGovHealthAlone%:16          HighMidWhite%: 6   HighMidBlack%:15
##  HighGovHealthAlone%   :25          HighWhite%   : 7   HighBlack%   :19
##
##
##         f.pctasian           f.pctotherrace     f.pctmarriedhouseholds
##  LowAsian%    :23   LowOtherRace%    :10     LowMarried%    :39
##  LowMidAsian% : 8   LowMidOtherRace% :10     LowMidMarried% : 8
##  HighMidAsian%: 6   HighMidOtherRace%:12     HighMidMarried%: 8
##  HighAsian%   :37   HighOtherRace%   :42     HighMarried%   :19
##
##
##        f.birthrate univariate_outlier_count
##  LowBirth%    :18   Min.   :0.000
##  LowMidBirth% :20   1st Qu.:2.250
##  HighMidBirth%:11   Median :4.000
##  HighBirth%   :25   Mean   :4.541
##                     3rd Qu.:6.000
##                     Max.   :9.000
```

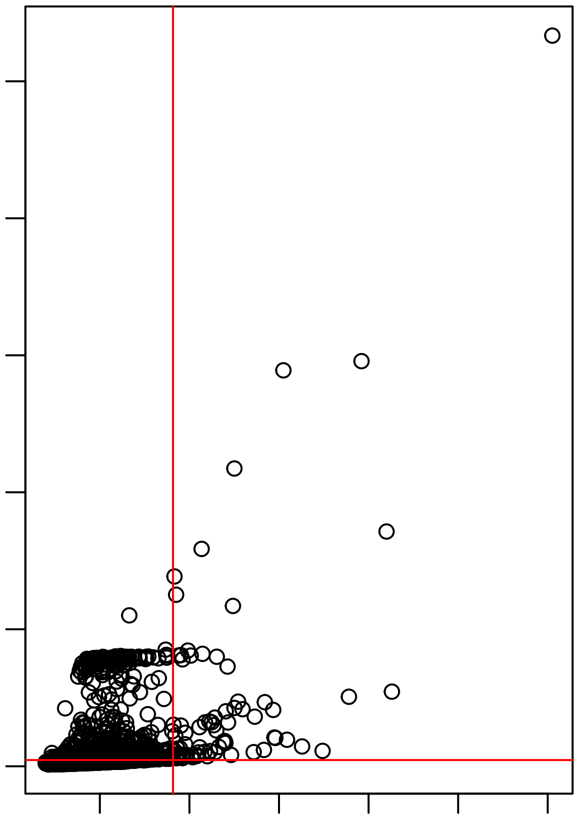
```
summary(df)
```

```
##   avganncount      avgdeathsperyear  target_deathrate incidencerate
##  Min.   :    7.0   Min.   :    3.0   Min.   : 59.7    Min.   : 201.3
##  1st Qu.:   79.5   1st Qu.:   29.0   1st Qu.:161.6    1st Qu.: 421.5
##  Median :  174.0   Median :   62.0   Median :178.4    Median : 453.5
##  Mean   :  611.0   Mean   :  187.2   Mean   :178.9    Mean   : 448.8
##  3rd Qu.:  495.5   3rd Qu.:  139.5   3rd Qu.:195.3    3rd Qu.: 481.3
##  Max.   :38150.0   Max.   :14010.0   Max.   :362.8    Max.   :1206.9
##    medincome         popest2015        povertypercent   studypercap
##  Min.   : 22640   Min.   :      829   Min.   : 3.70   Min.   :   0.00
##  1st Qu.: 39006   1st Qu.:    12106   1st Qu.:12.20   1st Qu.:   0.00
##  Median : 45439   Median :    27052   Median :15.70   Median :   0.00
##  Mean   : 47118   Mean   :   103705   Mean   :16.82   Mean   : 147.44
##  3rd Qu.: 52501   3rd Qu.:    65836   3rd Qu.:20.40   3rd Qu.:  73.56
##  Max.   :110507   Max.   :10170292   Max.   :44.00   Max.   :9762.31
##    binnedinc        medianage     medianagemale   medianagefemale
##  Min.   :28429   Min.   :23.30   Min.   :23.00   Min.   :23.9
##  1st Qu.:38888   1st Qu.:37.90   1st Qu.:36.40   1st Qu.:39.3
##  Median :46611   Median :40.90   Median :39.50   Median :42.4
##  Mean   :48942   Mean   :40.87   Mean   :39.61   Mean   :42.2
##  3rd Qu.:52796   3rd Qu.:43.90   3rd Qu.:42.60   3rd Qu.:45.3
##  Max.   :93565   Max.   :59.00   Max.   :60.20   Max.   :58.2
##   geography          percentmarried   pctnohs18_24     pcths18_24
##  Length:1823        Min.   :23.10   Min.   : 0.50   Min.   : 0.00
##  Class :character   1st Qu.:47.80   1st Qu.:13.00   1st Qu.:29.30
##  Mode  :character   Median :52.50   Median :17.20   Median :34.70
##                     Mean   :51.91   Mean   :18.31   Mean   :35.05
```

```
##                   3rd Qu.:56.40   3rd Qu.:22.75   3rd Qu.:40.50
##                   Max.   :68.00   Max.   :59.10   Max.   :72.50
##  pctbachdeg18_24   pcths25_over   pctbachdeg25_over pctemployed16_over
##  Min.   : 0.000   Min.   : 8.30   Min.   : 2.50     Min.   :23.90
##  1st Qu.: 3.200   1st Qu.:30.40   1st Qu.: 9.30     1st Qu.:48.60
##  Median : 5.400   Median :35.30   Median :12.30     Median :54.40
##  Mean   : 6.172   Mean   :34.81   Mean   :13.23     Mean   :54.17
##  3rd Qu.: 8.100   3rd Qu.:39.70   3rd Qu.:15.90     3rd Qu.:60.10
##  Max.   :51.800   Max.   :52.70   Max.   :42.20     Max.   :80.10
##  pctunemployed16_over pctprivatecoverage pctempprivcoverage pctpubliccoverage
##  Min.   : 0.700       Min.   :23.40      Min.   :14.30      Min.   :11.20
##  1st Qu.: 5.500       1st Qu.:57.50      1st Qu.:34.60      1st Qu.:30.95
##  Median : 7.500       Median :65.10      Median :41.10      Median :36.30
##  Mean   : 7.865       Mean   :64.42      Mean   :41.22      Mean   :36.21
##  3rd Qu.: 9.800       3rd Qu.:72.05      3rd Qu.:47.65      3rd Qu.:41.40
##  Max.   :29.400       Max.   :89.60      Max.   :70.20      Max.   :62.70
##  pctpubliccoveragealone    pctwhite        pctblack          pctasian
##  Min.   : 2.60          Min.   :12.27   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.:14.95          1st Qu.:77.63   1st Qu.: 0.6369   1st Qu.: 0.2566
##  Median :18.70          Median :90.06   Median : 2.2965   Median : 0.5460
##  Mean   :19.18          Mean   :83.93   Mean   : 9.0686   Mean   : 1.2175
##  3rd Qu.:23.00          3rd Qu.:95.58   3rd Qu.:10.8201   3rd Qu.: 1.2398
##  Max.   :46.60          Max.   :99.69   Max.   :85.9478   Max.   :37.1569
##   pctotherrace     pctmarriedhouseholds   birthrate              f.avganncount
##  Min.   : 0.0000   Min.   :22.99        Min.   : 0.000   LowCaseCount    :460
##  1st Qu.: 0.2838   1st Qu.:47.85        1st Qu.: 4.525   LowMidCaseCount :457
##  Median : 0.7779   Median :51.72        Median : 5.355   HighMidCaseCount:455
##  Mean   : 1.9907   Mean   :51.40        Mean   : 5.600   HighCaseCount   :451
##  3rd Qu.: 2.0957   3rd Qu.:55.47        3rd Qu.: 6.415
##  Max.   :41.9303   Max.   :71.40        Max.   :21.326
##        f.avgdeathsperyear          f.deathrate          f.incidencerate
##  LowMortCount    :461      LowDeathrate    :453   LowDiagnPerCap    :456
##  LowMidMortCount :455      LowMidDeathrate :458   LowMidDiagnPerCap :408
##  HighMidMortCount:456      HighMidDeathrate:456   HighMidDiagnPerCap:503
##  HighMortCount   :451      HighDeathrate   :456   HighDiagnPerCap   :456
##
##
##           f.medincome      f.popest2015    f.povertypercent    f.studypercap
##  LowMedianInc    :458   LowPop    :458   LowPov%    :452     NoTrials  :1161
##  LowMidMedianInc :457   LowMidPop :457   LowMidPov% :468     MidTrials : 333
##  HighMidMedianInc:456   HighMidPop:457   HighMidPov%:450     HighTrials: 329
##  HighMedianInc   :452   HighPop   :451   HighPov%   :453
##
##
##           f.binnedinc      f.medianage        f.medianagemale
##  LowIncPerCap    :366   LowAge    :452   LowAgeMale    :459
##  LowMidIncPerCap :529   LowMidAge :464   LowMidAgeMale :469
##  HighMidIncPerCap:558   HighMidAge:460   HighMidAgeMale:446
##  HighIncPerCap   :370   HighAge   :447   HighAgeMale   :449
##
##
##        f.medianagefemale    state                  f.percentmarried
##  LowAgeFemale    :454     Length:1823        LowMarriage%    :458
##  LowMidAgeFemale :469     Class :character   LowMidMarriage% :458
```

```
## HighMidAgeFemale:448    Mode  :character   HighMidMarriage%:453
## HighAgeFemale   :452                       HighMarriage%   :454
##
##
##          f.pctnohs18_24        f.pcths18_24          f.pcths25_over
## LowNoHighsc%    :454   LowHighsc%    :454   Low25Highsc%    :451
## LowMidNoHighsc% :459   LowMidHighsc% :462   LowMid25Highsc% :468
## HighMidNoHighsc%:454   HighMidHighsc%:456   HighMid25Highsc%:446
## HighNoHighsc%   :456   HighHighsc%   :451   High25Highsc%   :458
##
##
##   f.pctbachdeg25_over     f.pctemployed16_over f.pctunemployed16_over
## LowBach%    :459      LowEmploy%    :459       Min.   :1.000
## LowMidBach% :458      LowMidEmploy% :458       1st Qu.:1.000
## HighMidBach%:462      HighMidEmploy%:467       Median :2.000
## HighBach%   :444      HighEmploy%   :439       Mean   :2.494
##                                                3rd Qu.:4.000
##                                                Max.   :4.000
##      f.pcuntemployed16_over      f.pctprivatecoverage
## LowUnEmploy%    :464        LowPrivate%    :459
## LowMidUnEmploy% :451        LowMidPrivate% :464
## HighMidUnEmploy%:451        HighMidPrivate%:451
## HighUnEmploy%   :457        HighPrivate%   :449
##
##
##            f.pctempprivcoverage        f.pctpubliccoverage
## LowEmployeeHealth%    :465       LowGovHealth%    :456
## LowMidEmployeeHealth% :453       LowMidGovHealth% :459
## HighMidEmployeeHealth%:456       HighMidGovHealth%:453
## HighEmployeeHealth%   :449       HighGovHealth%   :455
##
##
##          f.pctpubliccoveragealone         f.pctwhite          f.pctblack
## LowGovHealthAlone%    :456         LowWhite%    :452   LowBlack%    :458
## LowMidGovHealthAlone% :463         LowMidWhite% :457   LowMidBlack% :459
## HighMidGovHealthAlone%:455         HighMidWhite%:456   HighMidBlack%:451
## HighGovHealthAlone%   :449         HighWhite%   :458   HighBlack%   :455
##
##
##        f.pctasian            f.pctotherrace     f.pctmarriedhouseholds
## LowAsian%    :458   LowOtherRace%    :458    LowMarried%    :455
## LowMidAsian% :457   LowMidOtherRace% :457    LowMidMarried% :460
## HighMidAsian%:458   HighMidOtherRace%:455    HighMidMarried%:453
## HighAsian%   :450   HighOtherRace%   :453    HighMarried%   :455
##
##
##       f.birthrate  univariate_outlier_count
## LowBirth%    :457   Min.   :0.000
## LowMidBirth% :455   1st Qu.:0.000
## HighMidBirth%:452   Median :1.000
## HighBirth%   :459   Mean   :1.495
##                     3rd Qu.:2.000
##                     Max.   :9.000
```

```
df = df[-which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff)),]
```



Profiling