Explorative Datenanalyse

Kacper Bohaczyk

2024-12-22

Infos `state`

Das Dataset zeigt eine Auswahl an Daten über die 50 Staaten der USA an.
Bevölkerung in 100 Tausenden,
Einkommen pro Person in USD,
Analphabetenrate als Prozent der Bevölkerung,
Lebenserwartung in Jahren,
Mord und Totschlag pro 100 Tausend Einwohner,
Menschen mit High-School abschlüssen als Prozent der Bevölkerung,
Tage mit Temperatur unter 0° C in großen Städten (1931-1960),
Fläche der Staaten in Quadratmeilen.

```
?state.x77
summary(state.x77)
```

```
##    Population        Income       Illiteracy       Life Exp
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
##  Median : 2838   Median :4519   Median :0.950   Median :70.67
##  Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88
##  3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89
##  Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60
##      Murder          HS Grad          Frost             Area
##  Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049
##  1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
##  Median : 6.850   Median :53.25   Median :114.50   Median : 54277
##  Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
##  3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81162
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
```

Bevölkerung

Median: 2838.5
Mitellwert: 4246.42
Standardabweichung: 4464.49
Quartildistanz: 3889

Es gibt einen großen unterschied zwishcen den Median und dem Mittlewert. Dadurch erkennt man das die Ränder stark verzehren. Für uns ist der Median mehr sinnvoll

```r
data(state)
round(median(state.x77[,"Population"]),2)
```

```
## [1] 2838.5
```

```r
round(mean(state.x77[,"Population"]),2)
```

```
## [1] 4246.42
```

```r
round(sd(state.x77[,"Population"]),2)
```

```
## [1] 4464.49
```

```r
round(IQR(state.x77[,"Population"]),2)
```

```
## [1] 3889
```

Einkommen pro Person

Median: 4519
Mitellwert: 4435.8
Standardabweichung: 614.47
Quartildistanz: 820.75

Der Unterschied zwischen dem Median und dem Mittlewert ist nicht sehr groß Die Daten werden durch die relativen großen Ausreißer am Rande des Sets verzogen.

```r
round(median(state.x77[,"Income"]),2)
```

```
## [1] 4519
```

```r
round(mean(state.x77[,"Income"]),2)
```

```
## [1] 4435.8
```

```r
round(sd(state.x77[,"Income"]),2)
```

```
## [1] 614.47
```

```r
round(IQR(state.x77[,"Income"]),2)
```

```
## [1] 820.75
```

Analphabetenrate

Median: 0.95
Mitellwert: 1.17
Standardabweichung: 0.61
Quartildistanz: 0.95

Es gibt keine großen Ausreißer. Die Daten sind ähnlich

```
round(median(state.x77[,"Illiteracy"]),2)
```

```
## [1] 0.95
```

```
round(mean(state.x77[,"Illiteracy"]),2)
```

```
## [1] 1.17
```

```
round(sd(state.x77[,"Illiteracy"]),2)
```

```
## [1] 0.61
```

```
round(IQR(state.x77[,"Illiteracy"]),2)
```

```
## [1] 0.95
```

Lebenserwartung in Jahren

Median: 70.67
Mitellwert: 70.88
Standardabweichung: 1.34
Quartildistanz: 1.78

Erstaunlicherweise gibt es keine großen Ausreißer.

```
round(median(state.x77[,"Life Exp"]),2)
```

```
## [1] 70.67
```

```
round(mean(state.x77[,"Life Exp"]),2)
```

```
## [1] 70.88
```

```
round(sd(state.x77[,"Life Exp"]),2)
```

```
## [1] 1.34
```

```r
round(IQR(state.x77[,"Life Exp"]),2)
```

```
## [1] 1.78
```

Mord und Totschlag pro 100 Tausend Einwohner

Median: 6.85
Mitellwert: 7.38
Standardabweichung: 3.69
Quartildistanz: 6.32

Keine großen Ausreißer

```r
round(median(state.x77[,"Murder"]),2)
```

```
## [1] 6.85
```

```r
round(mean(state.x77[,"Murder"]),2)
```

```
## [1] 7.38
```

```r
round(sd(state.x77[,"Murder"]),2)
```

```
## [1] 3.69
```

```r
round(IQR(state.x77[,"Murder"]),2)
```

```
## [1] 6.32
```

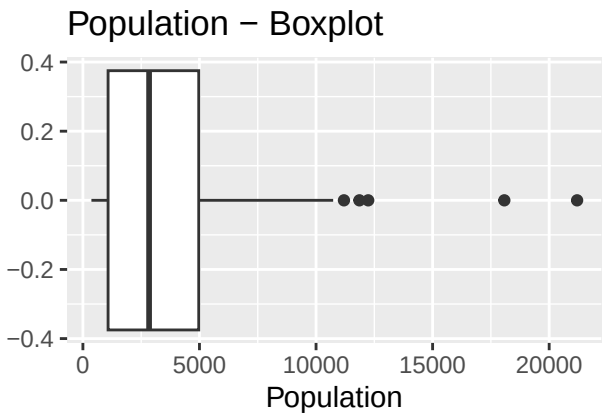
Plotten

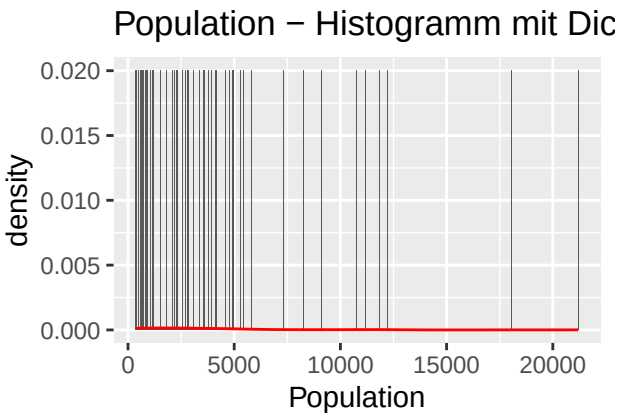
```r
library(ggplot2)
library(gridExtra)
# Daten laden
data(state)
df <- as.data.frame(state.x77)

# Variablennamen für die Grafiken anpassen
names(df)[4] <- "LifeExp"  # Ändern von "Life Exp" zu "LifeExp"

# Variablen auswählen
vars <- c("Population", "Income", "Illiteracy", "LifeExp", "Murder")

# Funktion, um die vier Grafiktypen für eine gegebene Variable zu erstellen
plot_matrix <- function(df, var) {
  # Boxplot
```

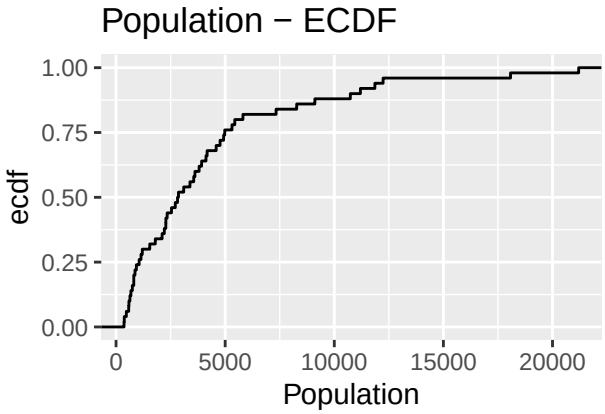
```r
  p1 <- ggplot(df, aes_string(x = var)) +
    geom_boxplot() +
    ggtitle(paste(var, "- Boxplot"))

  # Histogramm mit Dichteschätzung
  p2 <- ggplot(df, aes_string(x = var)) +
    geom_histogram(aes(y = ..density..), binwidth = 1) +
    geom_density(col = "red") +
    ggtitle(paste(var, "- Histogramm mit Dichteschätzung"))

  # ECDF
  p3 <- ggplot(df, aes_string(x = var)) +
    stat_ecdf(geom = "step") +
    ggtitle(paste(var, "- ECDF"))

  # QQ-Plot
  p4 <- ggplot(df, aes_string(sample = var)) +
    stat_qq() +
    stat_qq_line() +
    ggtitle(paste(var, "- QQ-Plot"))

  # Anordnen der Plots in einer 2x2-Matrix
  grid.arrange(p1, p2, p3, p4, nrow = 2)
}

for (var in vars) {
  plot_matrix(df, var)
}
```

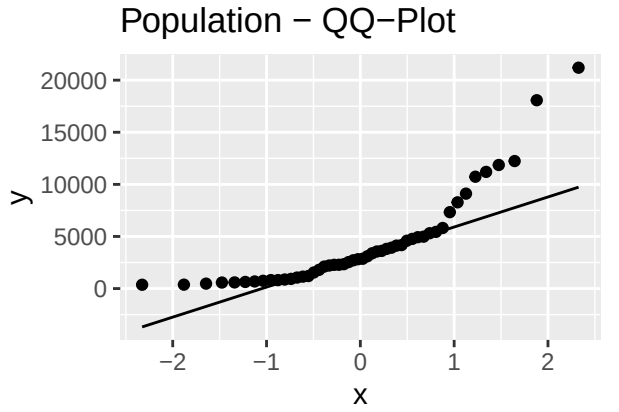
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

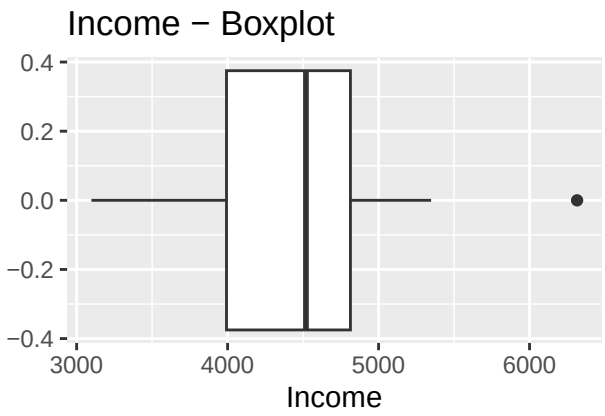
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

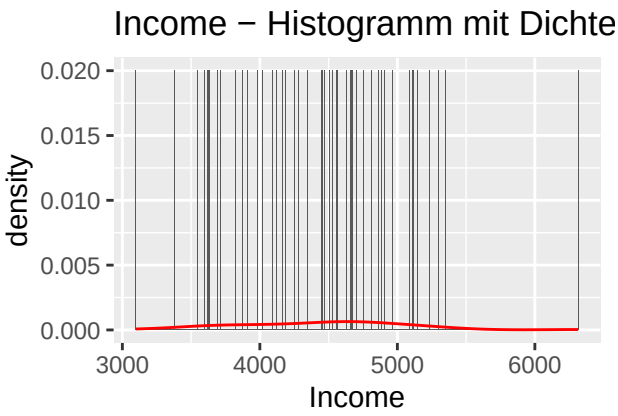

Population − Boxplot
0.4
0.2
0.0
−0.2
−0.4
0
5000
10000
15000
20000
Population

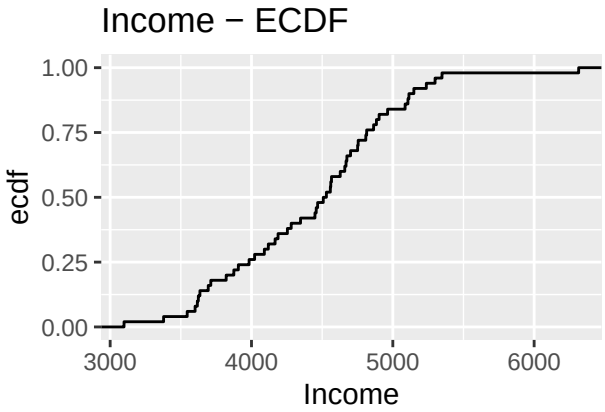

Population − Histogramm mit Dic
0.020
0.015
0.010
0.005
0.000
density
0
5000
10000
15000
20000
Population

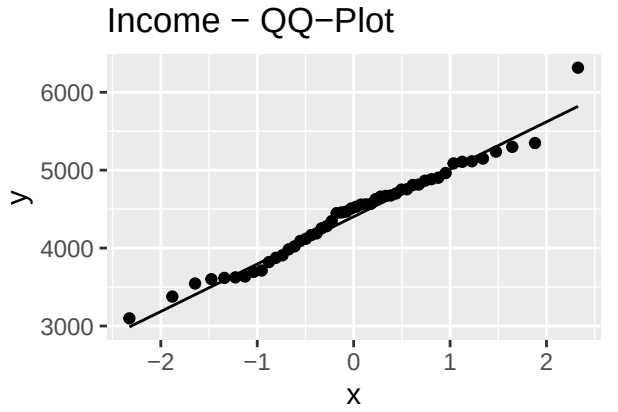

Population − ECDF
1.00
0.75
0.50
0.25
0.00
ecdf
0
5000
10000
15000
20000
Population

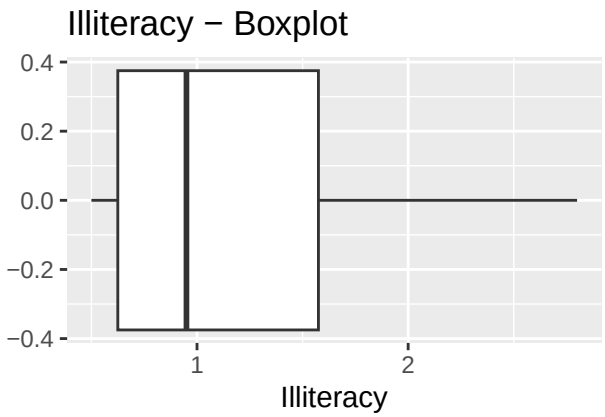

Population − QQ−Plot
20000
15000
10000
5000
0
y
−2
−1
0
1
2
x

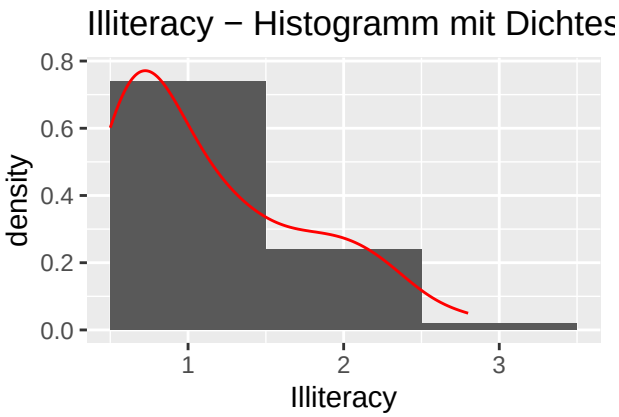

Income − Boxplot
0.4
0.2
0.0
−0.2
−0.4
3000
4000
5000
6000
Income

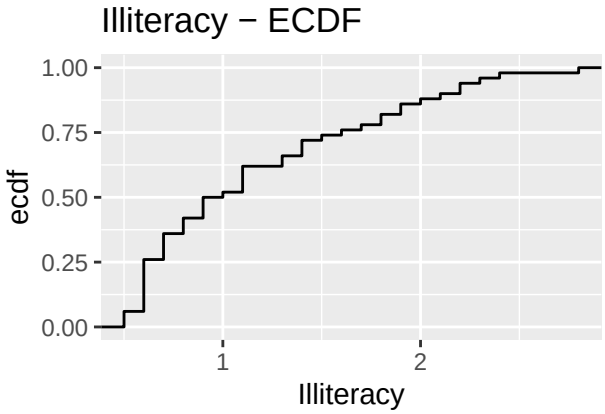

Income − Histogramm mit Dichte
0.020
0.015
0.010
0.005
0.000
density
3000
4000
5000
6000
Income

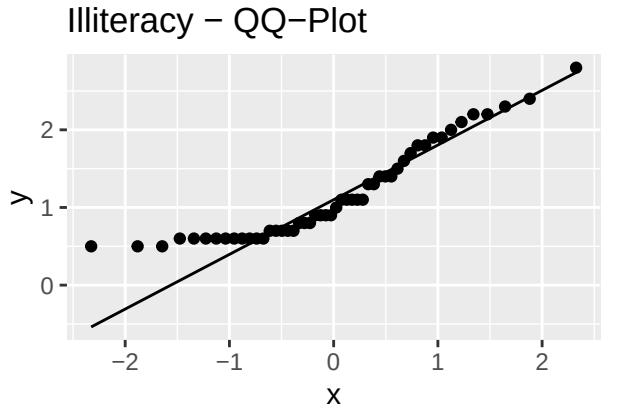

Income − ECDF
1.00
0.75
0.50
0.25
0.00
ecdf
3000
4000
5000
6000
Income

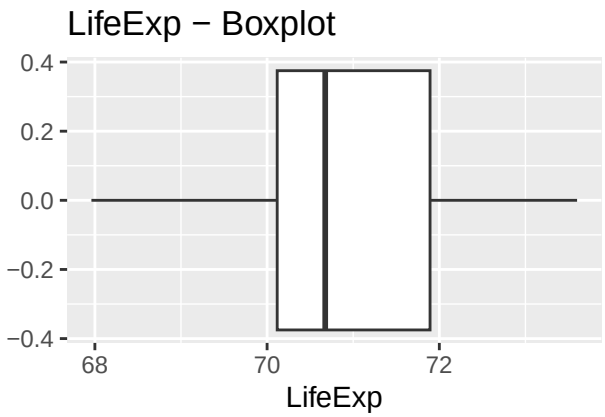

Income − QQ−Plot
6000
5000
4000
3000
y
−2
−1
0
1
2
x

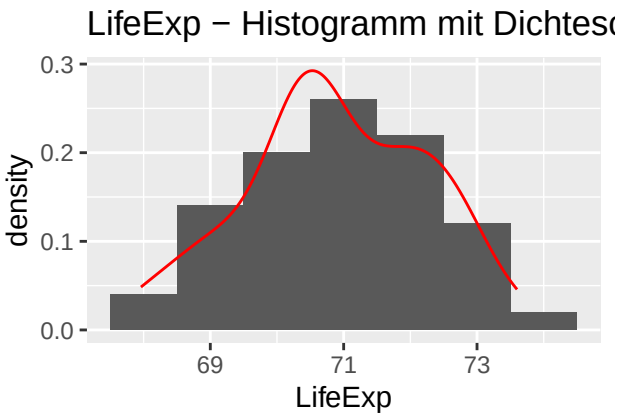

Illiteracy − Boxplot
0.4
0.2
0.0
−0.2
−0.4
1
2
Illiteracy

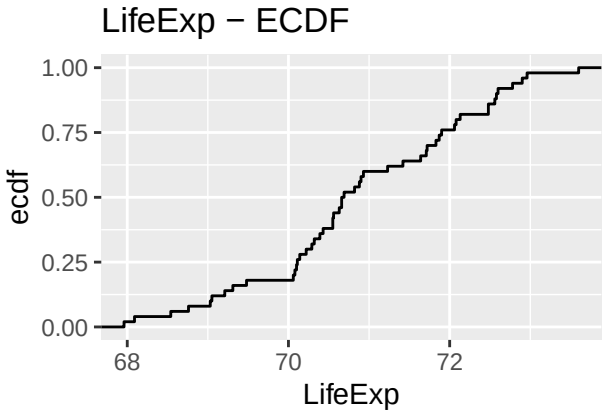

Illiteracy − Histogramm mit Dichtes
0.8
0.6
0.4
0.2
0.0
density
1
2
3
Illiteracy


Illiteracy − ECDF
1.00
0.75
0.50
0.25
0.00
ecdf
1
2
Illiteracy

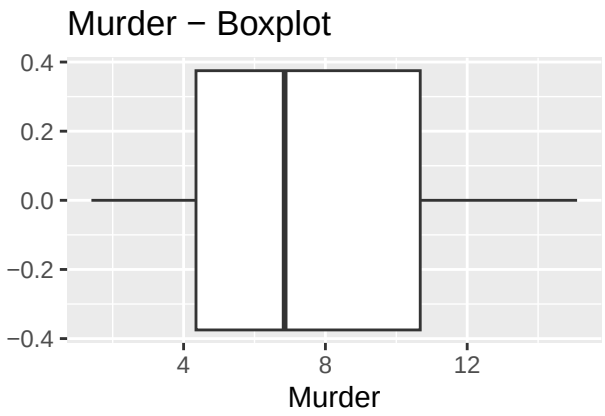

Illiteracy − QQ−Plot
2
1
0
y
−2
−1
0
1
2
x

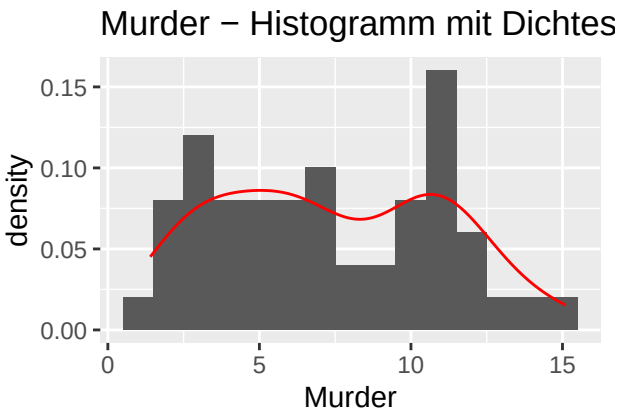

LifeExp − Boxplot
0.4
0.2
0.0
−0.2
−0.4
68
70
72
LifeExp

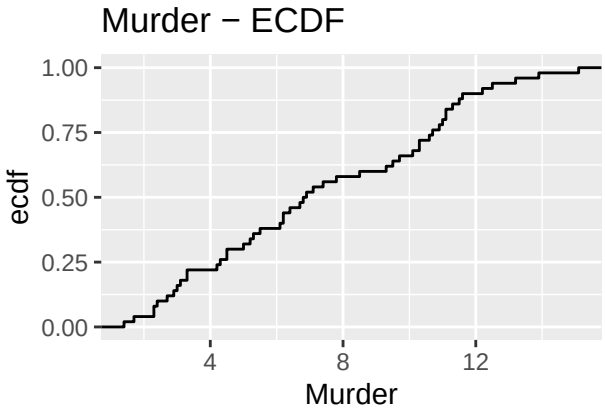

LifeExp − Histogramm mit Dichtes
0.3
0.2
0.1
0.0
density
69
71
73
LifeExp

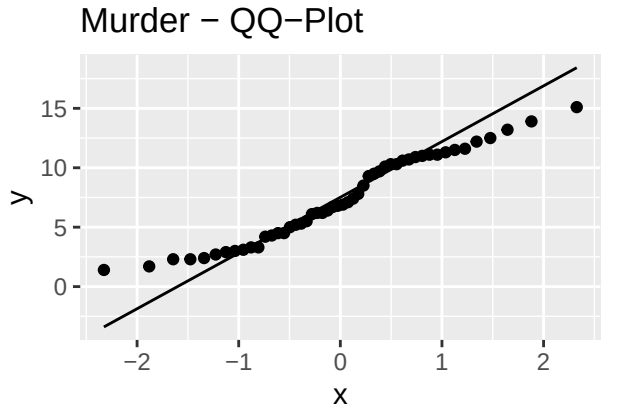

LifeExp − ECDF
1.00
0.75
0.50
0.25
0.00
ecdf
68
70
72
LifeExp


LifeExp − QQ−Plot
74
72
70
68
y
−2
−1
0
1
2
x









Ausreißer & Ränder Analyse

Die schlimmsten Ausreißer sind in 'Population', Städte wie Californien und New York haben beide ~20 Millionen einwohner, während Wyoming, Vermont und Alaska alle weniger als 500k haben.

Bei Einkommen gibt es auch Außreiser, da Missisipi, Vermont, Alabama, West Virgina, Kentucky und Louisiana alle unter 4000 DollerEinkommen haben und Alaska als einziger Staat über 6000$ Einkommen hat.

Die Analphabetenrate, hat keine Ausreißer aber starke Ränder.

Lebenserwartung hat weder starke Ränder oder Ausreißer, da es in allen Staaten sehr änhlich ist.

Mord hat auch keine Ausreißer, nur wieder starke Ränder.