Explorative Datenanalyse und Visualisierung

Zlabinger Christof

2024-01-12

Lade den Datensatz 'state.x77' in R. Beschreibe die Daten anhand der internen Hilfe.

```
##    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
```

Es handelt sich um eine Matrix mit 50 Zeilen und 8 Reihen welche Dated der US Staaten beinhalten. Jede Zeile entspricht einem Staat. Diese Reihen beinhalten die:

- Population im Jahr 1975 in 100 Einwohnern
- Das Einkommen pro Person in 1974
- Die Prozent an Analphabeten in 1970
- Die Lebenserwartung von 1969-1971
- Die Mordrate pro 100,000 Einwohnern
- Die Prozent der Highschool Absolventen
- Die Mittlere Anzahl an Tagen an denen es in der Hauptstadt oder in einer grossen Stadt, in den Jahren 1931-1960, es unter 0°c hatte.
- Die Flaeche der Laender

Ermittle mithilfe geeigneter Schätzer für die Lage (arithmetischer Mittelwert und Median sollen verglichen werden) und Streuung (Standardabweichung und Interquartilsdistanz sollen verglichen werden) der ersten 5 Variablen: Population, Income, Illiteracy, Life Exp(ectancy) und Murder.

Population

Median:

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

Mittelwert:

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

Aufgrund des Unterschiedes von Median und Mittelwert wird auf eine nicht Symetrische verteilung der Daten hingewiesen.

Interquartildistanz:

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

Standartabweichung:

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

Der Unterschied der IQR und der SD weisst auf nicht normalverteilte Werte oder aussreiserische Werte hin.

Income

Median:

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

Mittelwert:

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

Da der Unterschied zwischen Median und Mittelwert gering ist deuted dies auf eine annaehende Symetrie hin.

Interquartildistanz:

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

Standartabweichung:

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

Da die Differenz der IQR und SD nicht gross ist sind die Daten normalverteilt.

Illiteracy

Median:

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

Mittelwert:

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

Aufgrund des Unterschiedes von Median und Mittelwert wird auf eine nicht Symetrische verteilung der Daten hingewiesen.

Interquartildistanz:

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

Standartabweichung:

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

Der Unterschied der IQR und der SD weisst auf nicht normalverteilte Werte oder aussreiserische Werte hin.

Life Exp

Median:

```
## [1] 70.675
```

Mittelwert:

```
## [1] 70.8786
```

Da der Unterschied zwischen Median und Mittelwert sehr gering ist deuted dies auf eine Symetrie hin.

Interquartildistanz:

```
## [1] 1.775
```

Standartabweichung:

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

Der Unterschied der IQR und der SD weisst auf nicht normalverteilte Werte oder aussreiserische Werte hin.

Murder

Median:

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

Mittelwert:

```
## [1] 7.378
```

Aufgrund des Unterschiedes von Median und Mittelwert wird auf eine nicht Symetrische verteilung der Daten hingewiesen.

Interquartildistanz:

```
## [1] 6.325
```

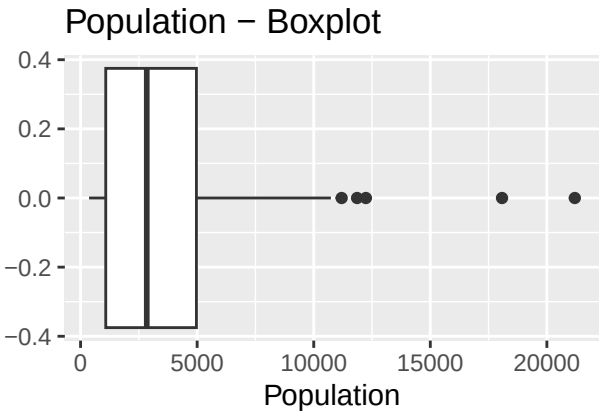
Standartabweichung:

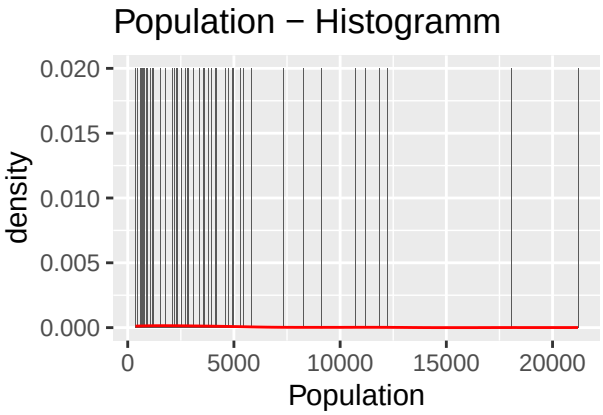
```
## [1] 3.69
```

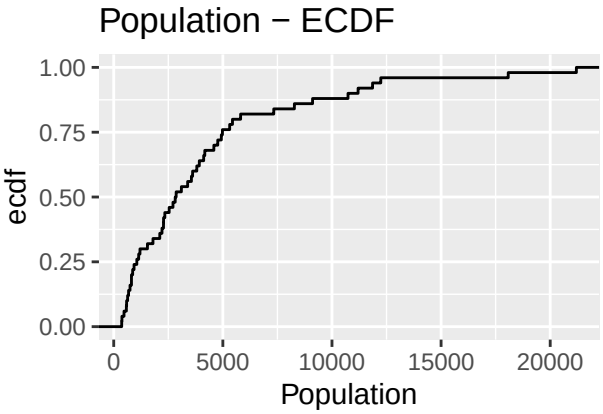
Der Unterschied der IQR und der SD weisst auf nicht normalverteilte Werte oder aussreiserische Werte hin.

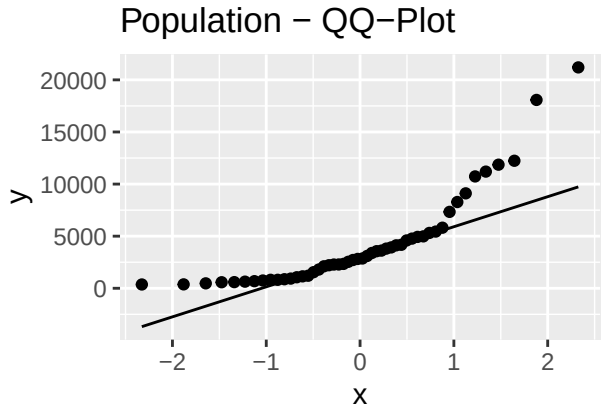
Stelle die Daten der ersten 5 Variablen, Population, Income, Illiteracy, Life Exp(ectancy) und Murder in geeigneter Weise graphisch dar, indem du Boxplot, Histogramm mit Dichteschätzung, ECDF und QQ-Plot verwendest.

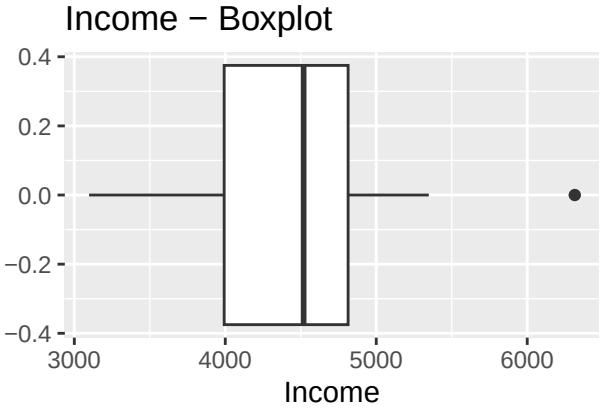
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this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

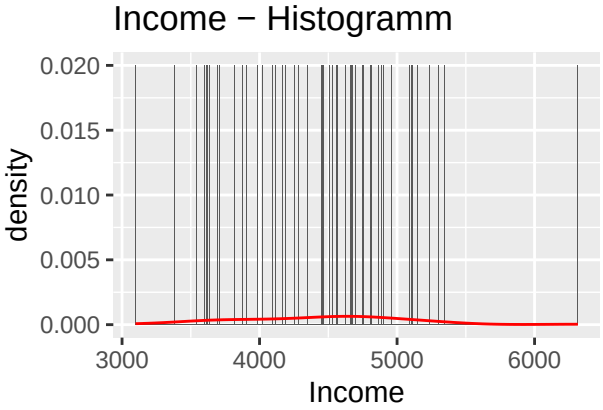

Population − Boxplot
0.4
0.2
0.0
−0.2
−0.4
0
5000
10000
15000
20000
Population

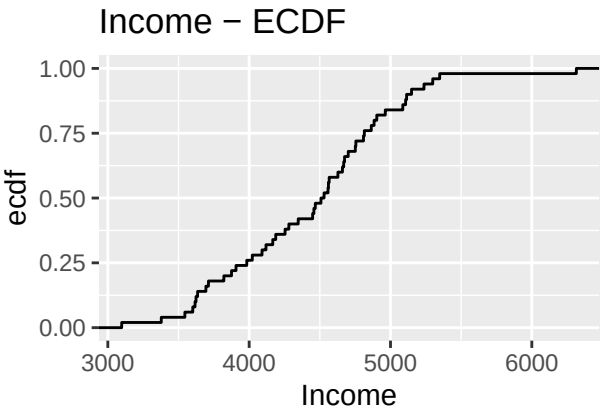

Population − Histogramm
0.020
0.015
0.010
0.005
0.000
density
0
5000
10000
15000
20000
Population

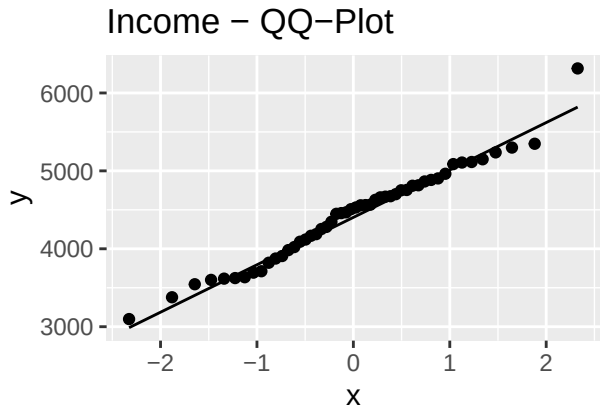

Population − ECDF
1.00
0.75
0.50
0.25
0.00
ecdf
0
5000
10000
15000
20000
Population

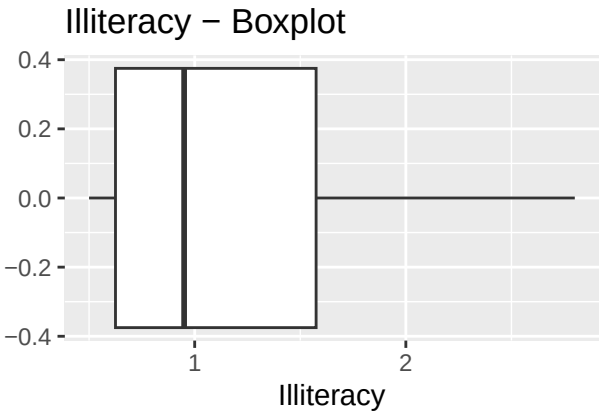

Population − QQ−Plot
20000
15000
10000
5000
0
y
−2
−1
0
1
2
x

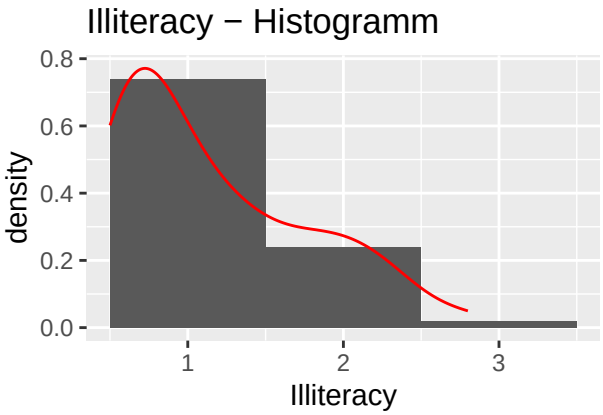

Income − Boxplot
0.4
0.2
0.0
−0.2
−0.4
3000
4000
5000
6000
Income

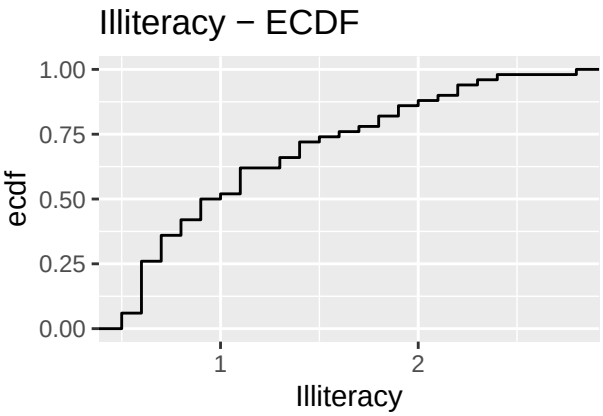

Income − Histogramm
0.020
0.015
0.010
0.005
0.000
density
3000
4000
5000
6000
Income

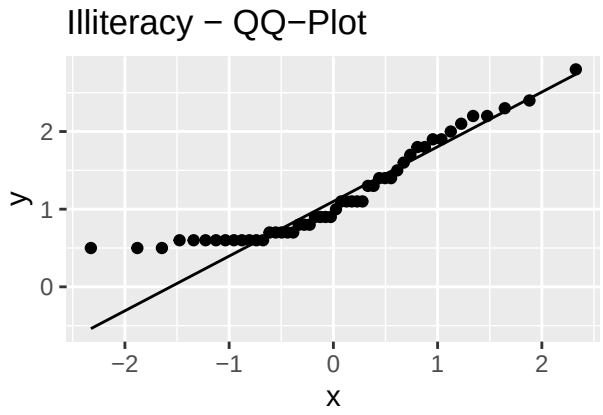

Income − ECDF
1.00
0.75
0.50
0.25
0.00
ecdf
3000
4000
5000
6000
Income

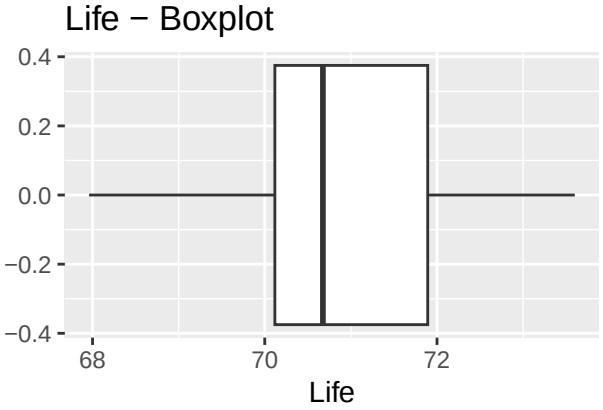

Income − QQ−Plot
6000
5000
4000
3000
y
−2
−1
0
1
2
x

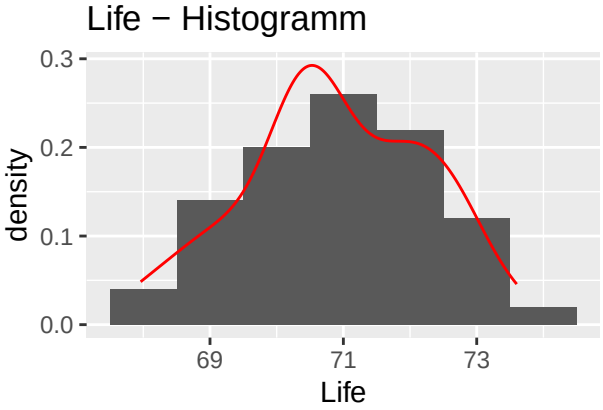

Illiteracy − Boxplot
0.4
0.2
0.0
−0.2
−0.4
1
2
Illiteracy

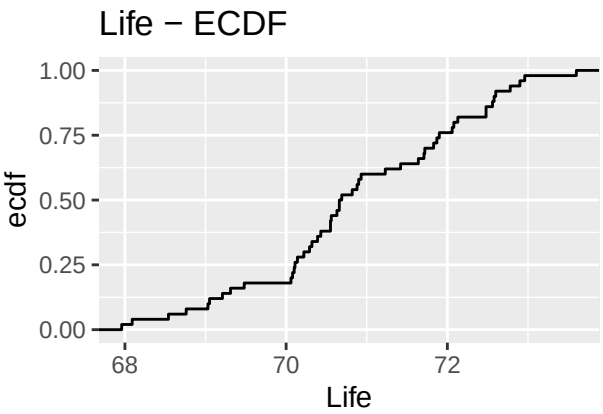

Illiteracy − Histogramm
0.8
0.6
0.4
0.2
0.0
density
1
2
3
Illiteracy

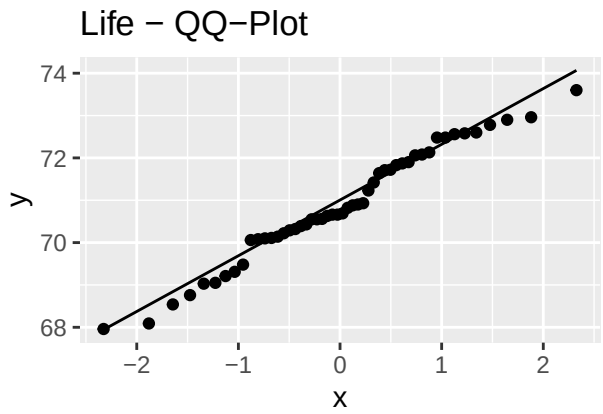

Illiteracy − ECDF
1.00
0.75
0.50
0.25
0.00
ecdf
1
2
Illiteracy

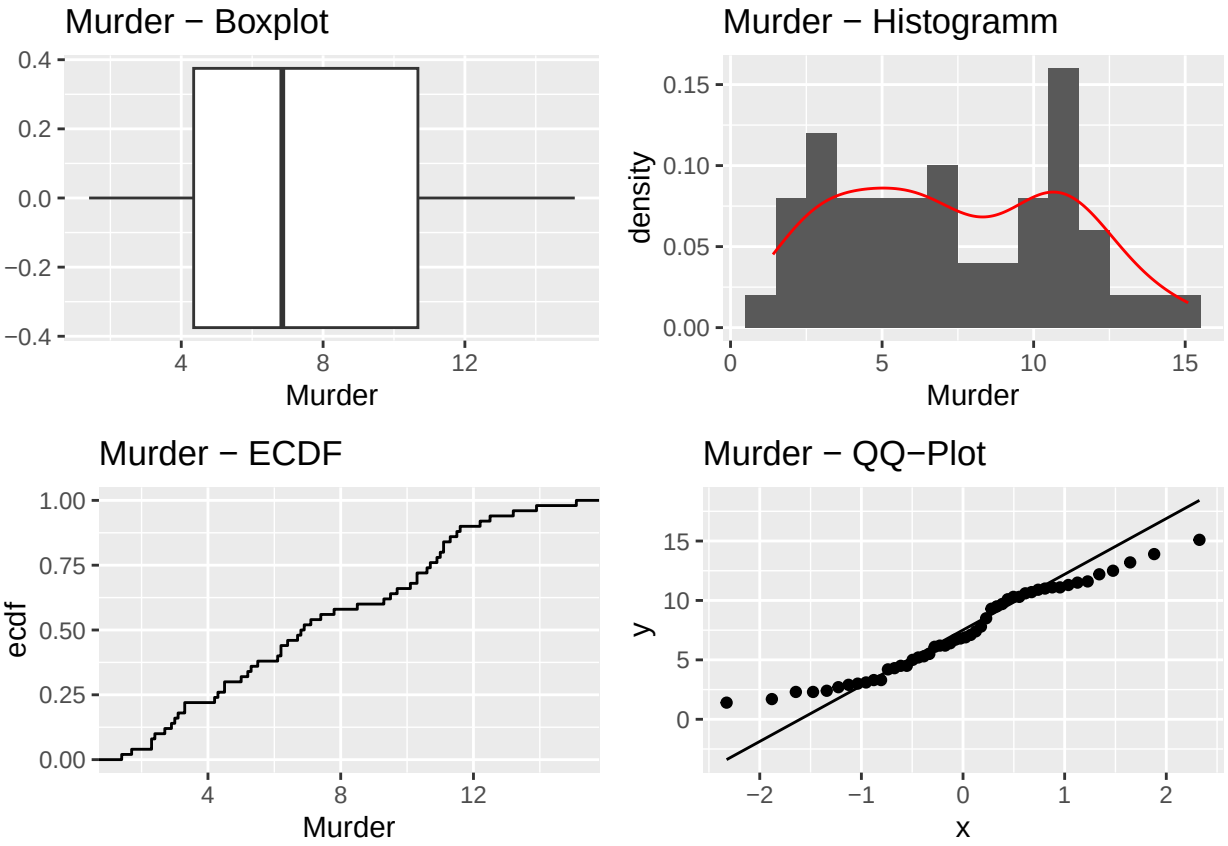

Illiteracy − QQ−Plot
2
1
0
y
−2
−1
0
1
2
x


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


Life − Histogramm
0.3
0.2
0.1
0.0
density
69
71
73
Life


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


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



Begründe anhand der graphischen Darstellung, ob es sich symmetrische oder schiefe Datenverteilungen handelt. Begründe anhand der graphischen Darstellungen, ob schwere oder leichte Ränder vorliegen (und auf welcher Seite).

Population

Die Population ist deutlich rechtsschief. Auf beiden seiten befindet sich ein schwerer Rand.

Income

Das Income ist leicht rechtsschief. Auf der linken Seite ein schwerer Rand und auf der rechten Seite ein leichter Rand.

Illiteracy

Die Illiteracy ist rechtsschief. Auf der linken Seite befindet sich ein leichter Rand und auf der rechten Seite ein schwerer Rand.

Life Exp.

Die Life Exp. ist symetrisch. Auf beiden Seiten ist ein leichter Rand vorhanden.

Murder

Bei der Murder rate ist eine symetrie zu erkennen. Auf beiden Seiten ist ein schwerer Rand zu erkennen.

Bestimme anhand graphischen Darstellungen aus Punkt 3. und der Erkenntnisse aus 4., ob Ausreißer vorliegen und welche Punkte dies sind.

Population

Bei der Population sind die schlimmsten Ausreisser:

- Alaska
- Wyoming
- Vermont

da sie alle unter 500k Einwohnern haben.

- California
- New York

da beide ca 20M Einwohner haben.

Income

Bei Income sind die schlimmsten Ausreisser:

- Mississippi
- Arkansas
- Louisiana

da das Einkommen unter 3.6k$ liegt.

- Alaska

da das Income ueber 6k$ liegt.

Illiteracy

Bei der Illiteracy gibt es keine ersichtlichen Ausreisser.

Life Exp.

Bei der Life Exp. gibt es auch keine ausreisserrischen Werte.

Murder

Bei der Murder rate gibt es ebenfalls keine Ausresser.

Beschreibung der Begriffe

- ordinale
  - Werte werden nach ihrer Reihenfolge geordnet
- nominal
  - Werte werden in Kategorien eingeteilt. Sie weisen keine Ordnung auf.
- metrisch rational
  - Hat einen logischen nullpunkt. Abstaende zwischen Werten ist gleich gross.
- metrisch intervallskaliert
  - Haben keinen logischen nullpunkt. Abstaende zwischen Werten ist gleich gross.
- absolute Häufigkeiten
  - Die tatsaechliche haeufigkeit der Werte.
- relative Häufigkeiten
  - Der anteil der absoluten Haufigkeit im Verhaeltniss zur Gesammtzahl.
- Mittelwert
  - Durchschnitt aller Ergebnissen.

- Median
  - Der Wert der genau bei 50% liegt.
- Varianz
  - Die Abweichung der Werteˆ2 im vergleich zum Mittelwert.
- Standardabweichung
  - Die durchschnittliche Abweichung vom Mittelwert.
- Interquartilsdistanz
  - Die Distanz zwischen dem 3. Quartil und dem 1. Quartil.
- Symmetrie
  - Die verteilung der Daten auf der linke und rechte Seite der Daten sind gleich.
- Schiefe
  - Auf welcher Seite die Daten eine asymetrische Verteilung aufweisen.
- schwere Ränder
  - Die verteilung der Daten an den Enden ist breiter als im normalfall.
- Ausreißer
  - Daten die von dem Rest der Daten abweichen.
- Histogramm
  - Zeigt die Haefigkeit der Daten in bestimmten Intervallen.
- Boxplot
  - Zeigt die Daten anhand einer Box und Whiskers an.
- QQ-Plot
  - Zeigt moegliche Abweichungen von der theoretischen Verteilung.