# Hausübung 2

Reutterer Maximilian, Sattler Lukas, Weinzierl Jakob

2023-11-26

**Aufgabe 1**

Explorieren und visualisieren Sie die Variablen "Fertility", "Agriculture", "Education", "Catholic" und "Infant. Mortality" aus dem R Datensatz swiss des R package utils. Betrachten Sie vorerst jede Variable als separate Stichprobe fur eindimensionale Exploration (ziehen Sie die Bedeutung der Variablen im Sachkontext in Betracht). Fur jede Variable: • wählen Sie sinnvolle Schätzer für Lokation, Variation, Schiefe and Gewicht in den Rändern. • Geben Sie dem Nutzer die Möglichkeit zwischen unterschiedlichen graphischen Darstellungen zu wechseln. Erklären Sie die Zusammenhänge und Eigenschaften der Daten, die sich aus diesen Visualisierungen erkennen lassen. – Sind die Daten symmetrisch/schief? – Haben die Daten schwere Ränder? – Bieten Sie robuste und nichtrobuste Lagemaße und Skalenmaße im Vergleich oder zur Auswahl an. – Sind die Daten (approximativ) normalverteilt? Was lässt sich über die Zusammenhänge zwischen den Variablen aussagen? (Tipp: scatterplot matrix.)

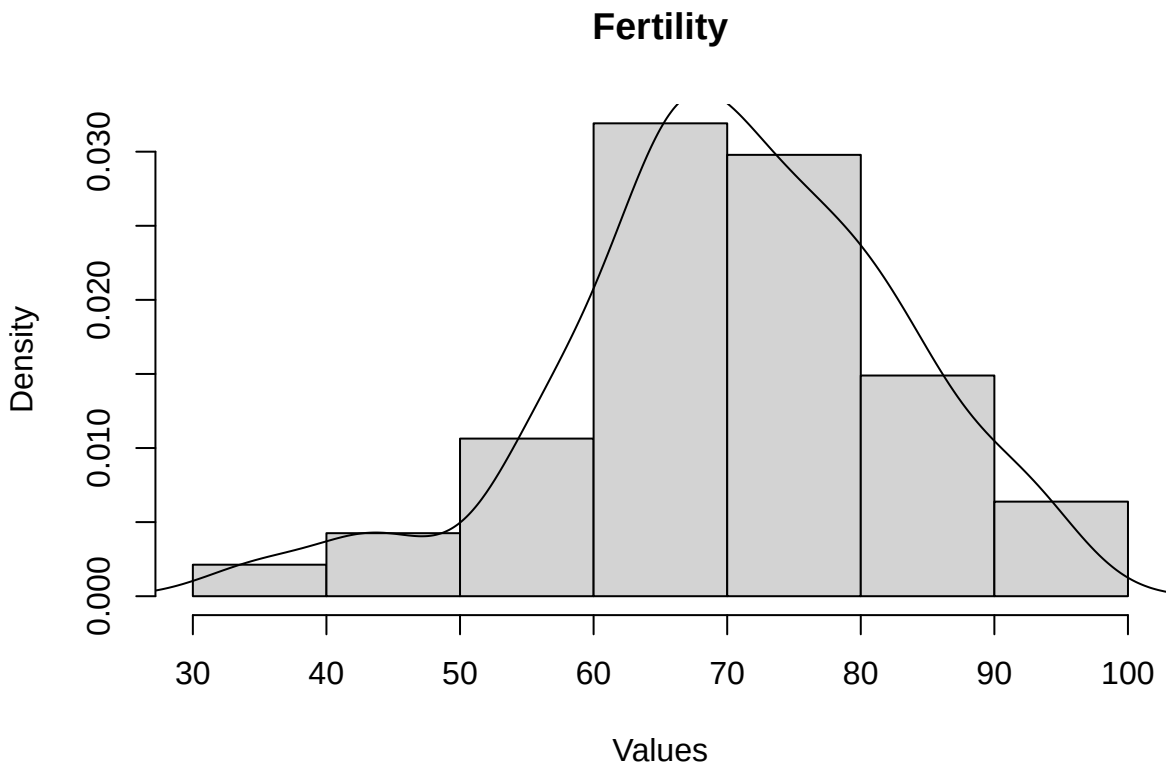
```
library("moments")
library("Hmisc")

data = data.frame(swiss)
summary(data)
```

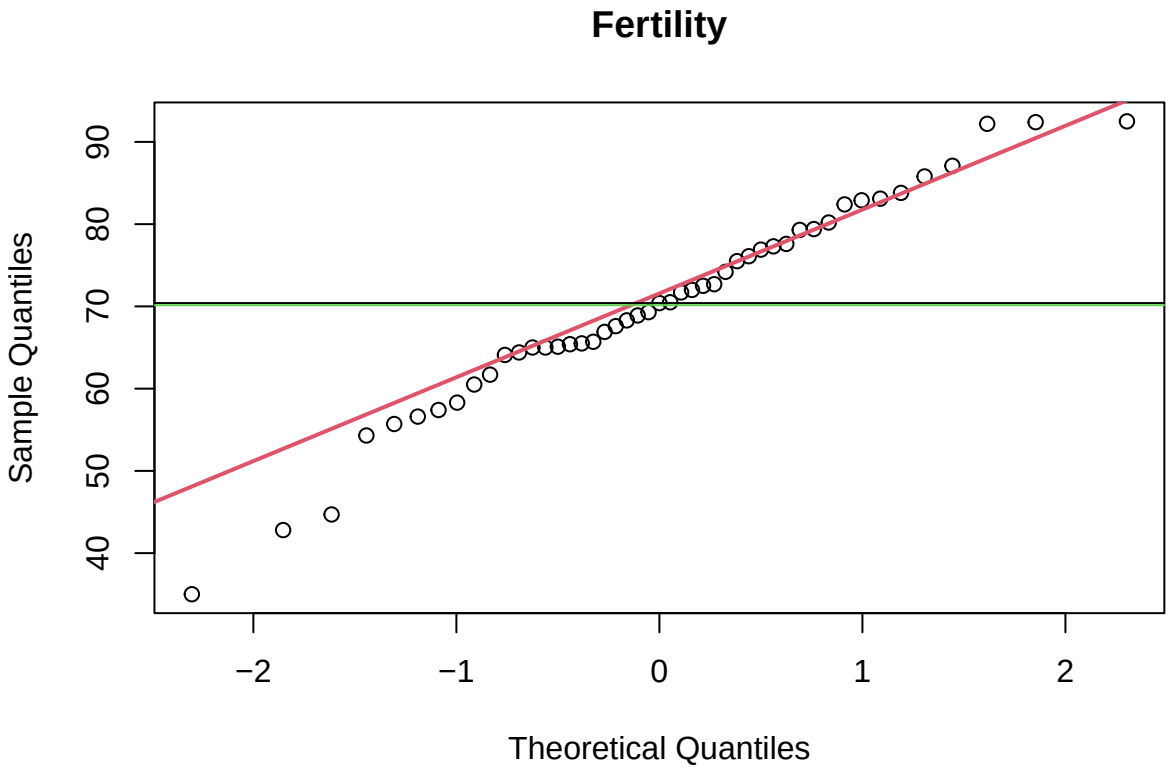
```
##    Fertility      Agriculture     Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##     Catholic       Infant.Mortality
##  Min.   :  2.150   Min.   :10.80
##  1st Qu.:  5.195   1st Qu.:18.15
##  Median : 15.140   Median :20.00
##  Mean   : 41.144   Mean   :19.94
##  3rd Qu.: 93.125   3rd Qu.:21.70
##  Max.   :100.000   Max.   :26.60
```

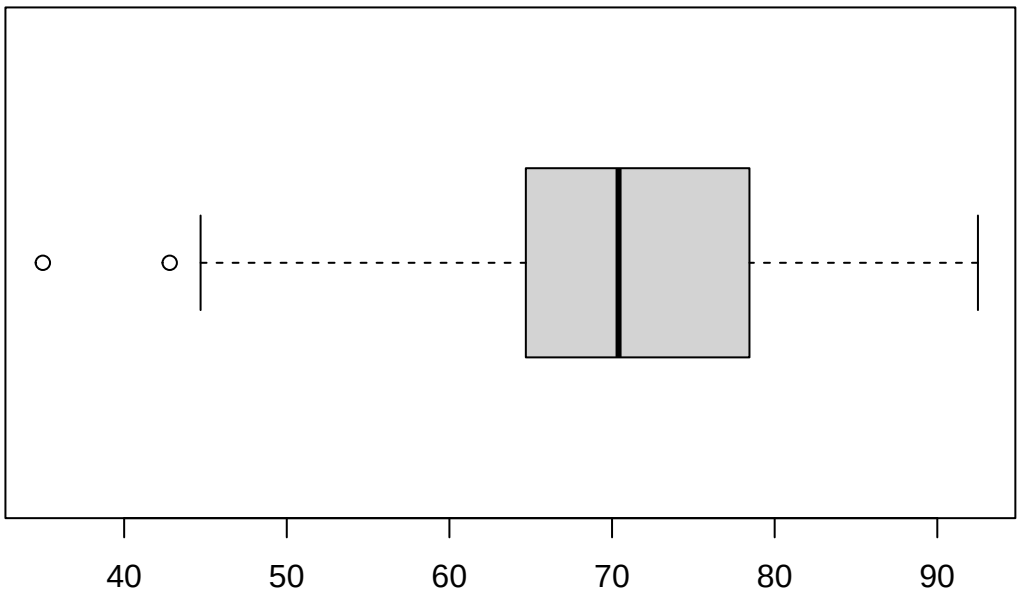
```
df <- swiss
i= 1
for (k in df){
  mycols <- colnames(df)

  i<-i+1
}
```

```
for (k in (1:6)) {

  print(mycols[k])
  print(shapiro.test(data[,k]))
  hist(data[,k],main = paste(mycols[k]),freq = F, xlab = "Values"); lines(density(data[,k]))
  qqnorm(data[,k], main = paste(mycols[k]))
  qqline(data[,k],col=2,lwd=2)
  abline(h=median(data[,k]))
  abline(h=mean(data[,k]),col=3)
  boxplot(data[,k], horizontal = T, main = paste(mycols[k]))
  print(skewness(data[,k]))
  print(kurtosis(data[,k]))
  print(sd(data[,k]))
}
```

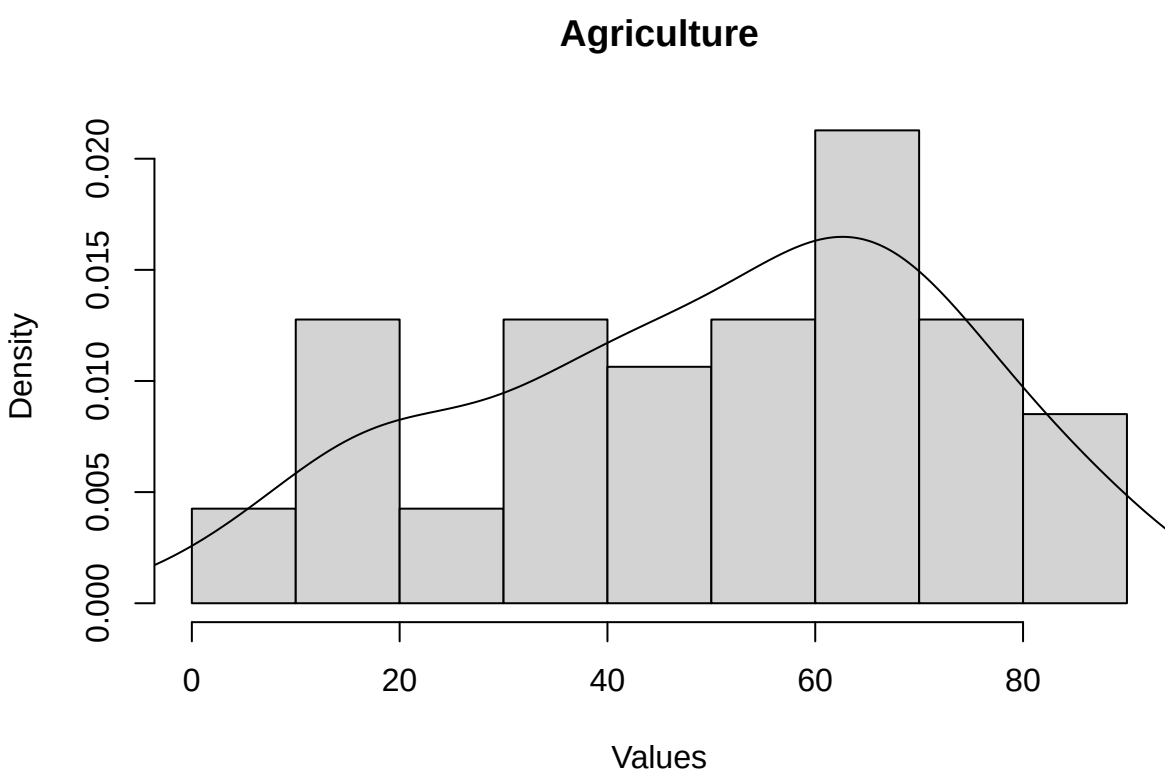
```
## [1] "Fertility"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.97307, p-value = 0.3449
```

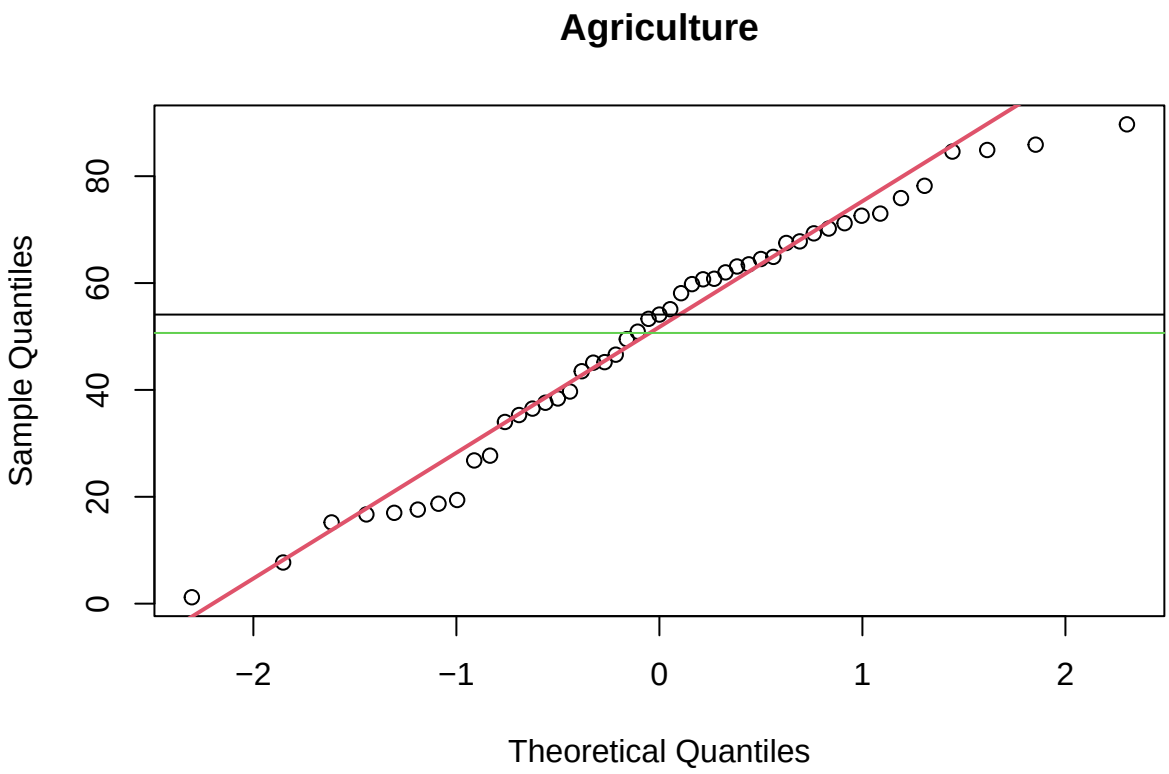


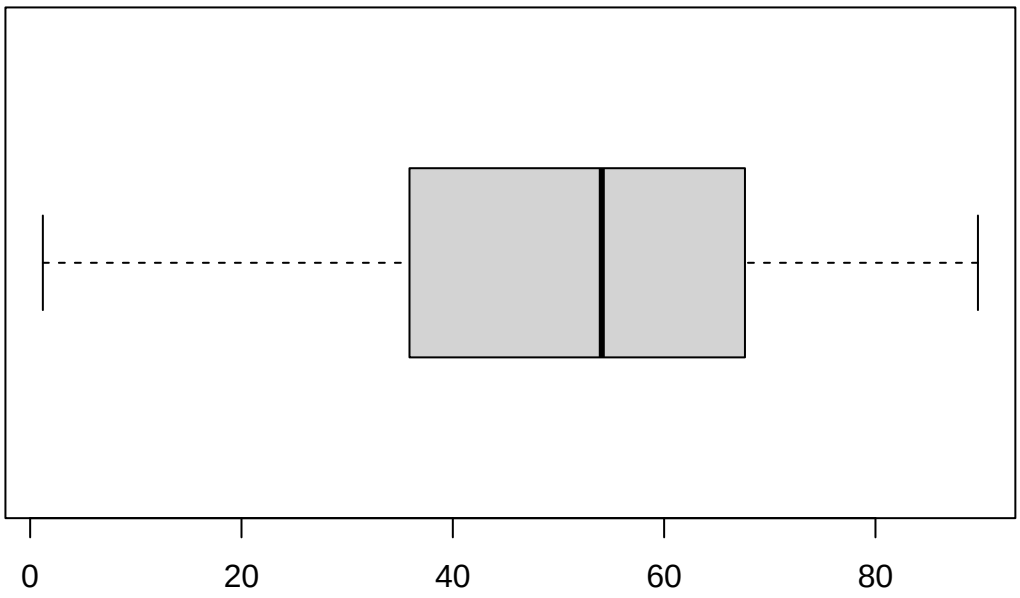

Fertility
Sample Quantiles
40
50
60
70
80
90
−2
−1
0
1
2
Theoretical Quantiles



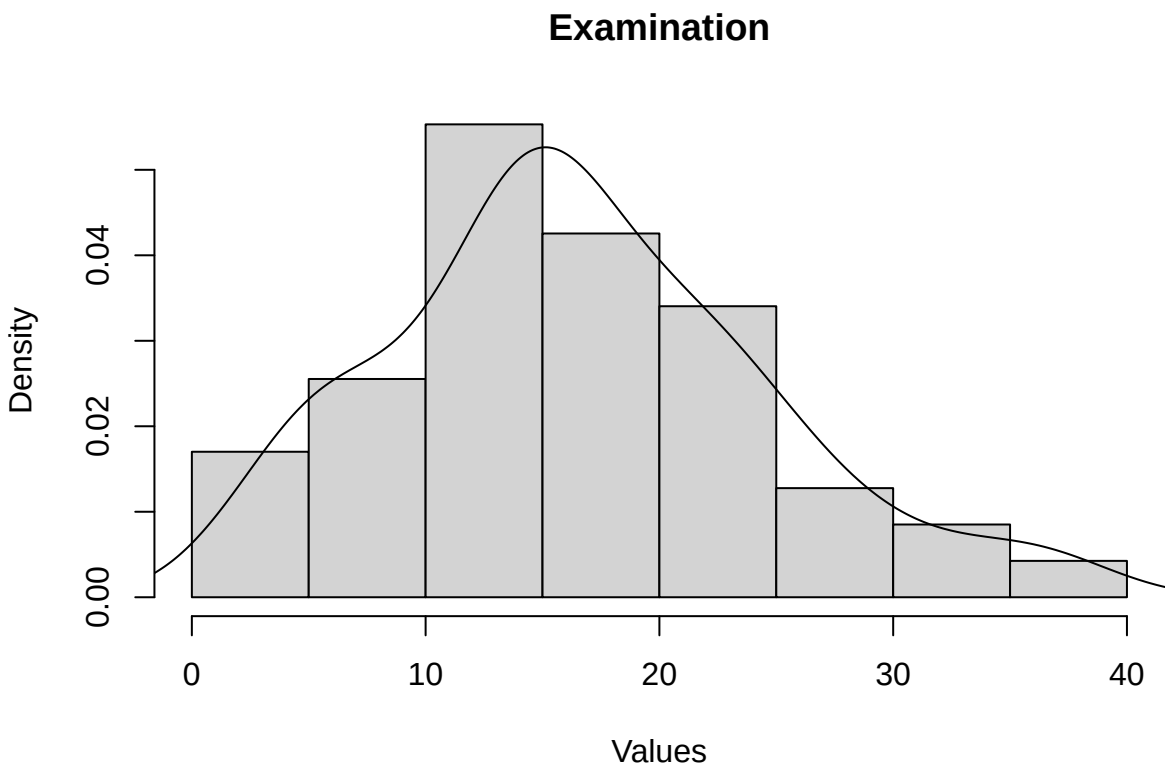

```
## [1] -0.4706269
## [1] 3.403232
## [1] 12.4917
## [1] "Agriculture"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.96643, p-value = 0.193
```

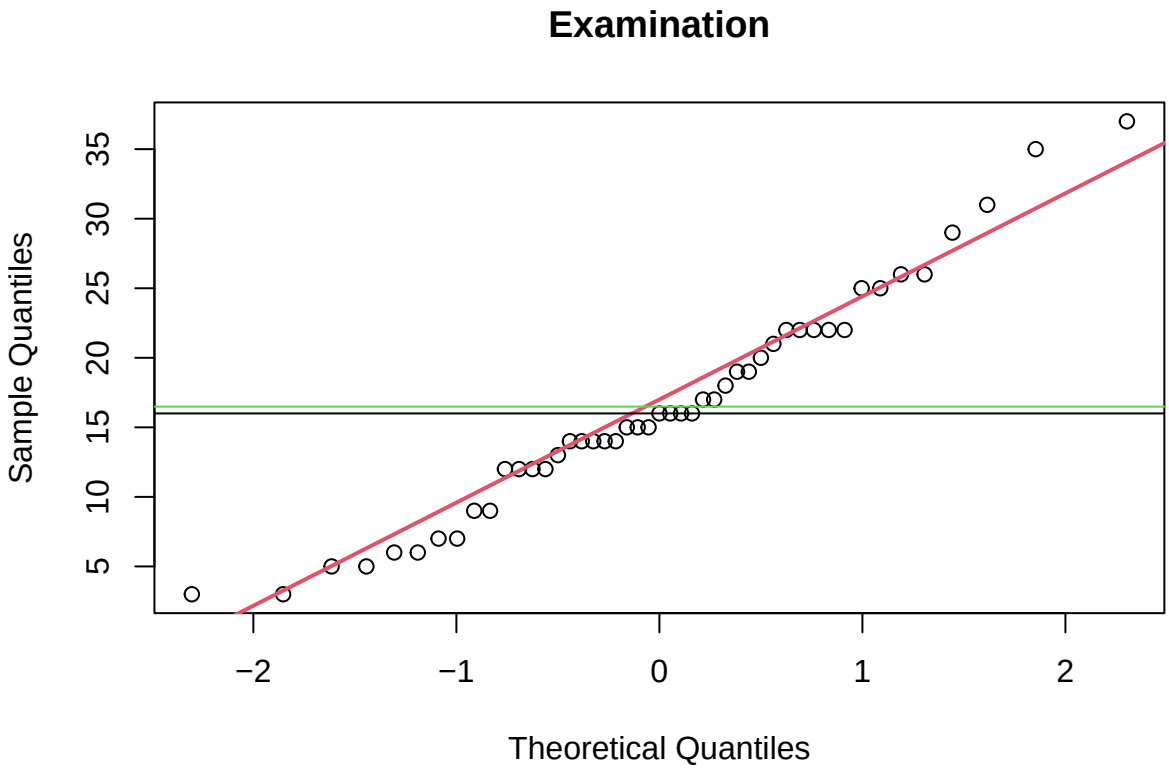

Agriculture
Density
0.000
0.005
0.010
0.015
0.020
0
20
40
60
80
Values

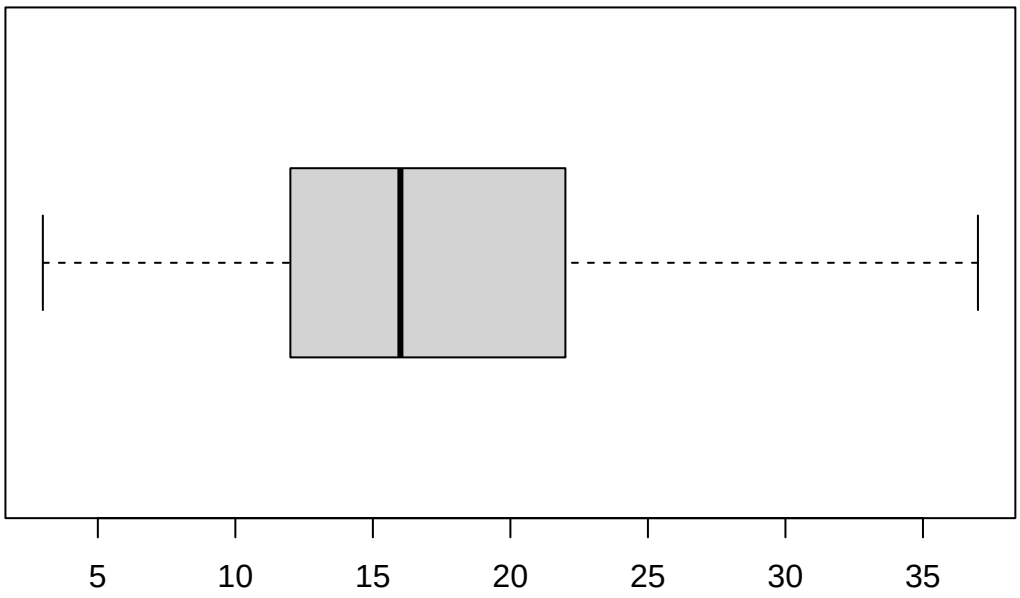

Agriculture
Sample Quantiles
0
20
40
60
80
−2
−1
0
1
2
Theoretical Quantiles



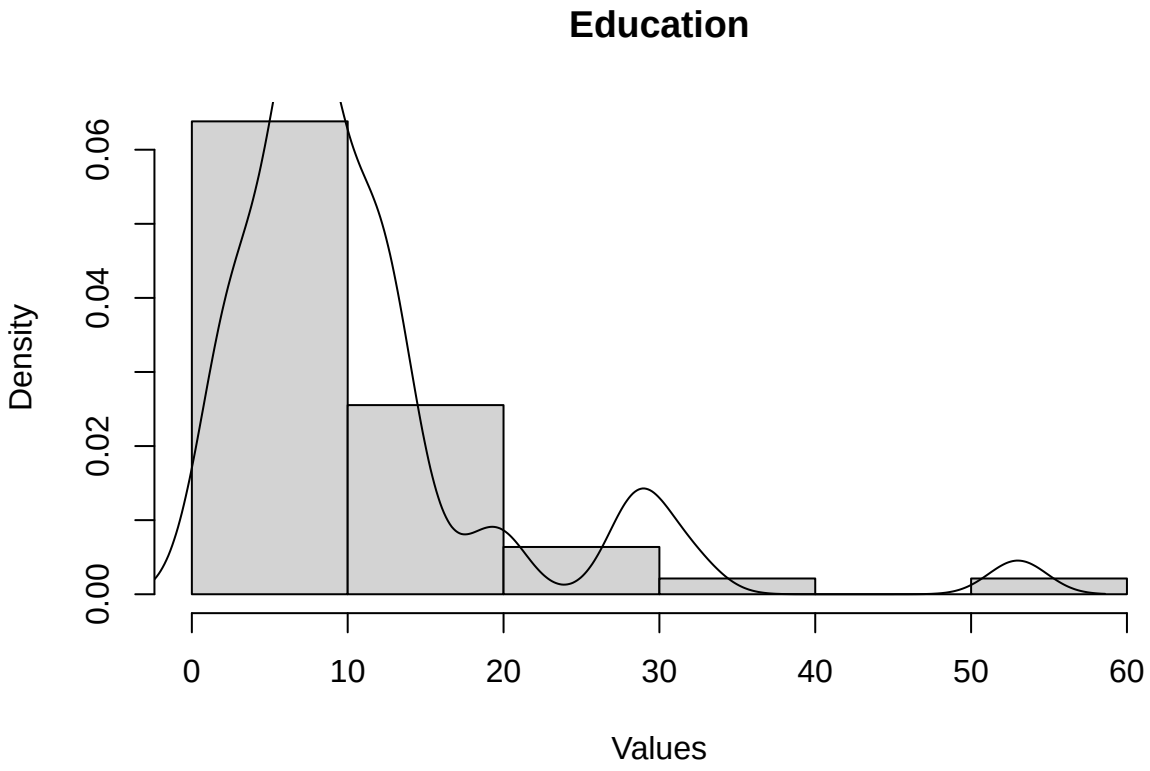

```
## [1] -0.330867
## [1] 2.207406
## [1] 22.71122
## [1] "Examination"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.96962, p-value = 0.2563
```

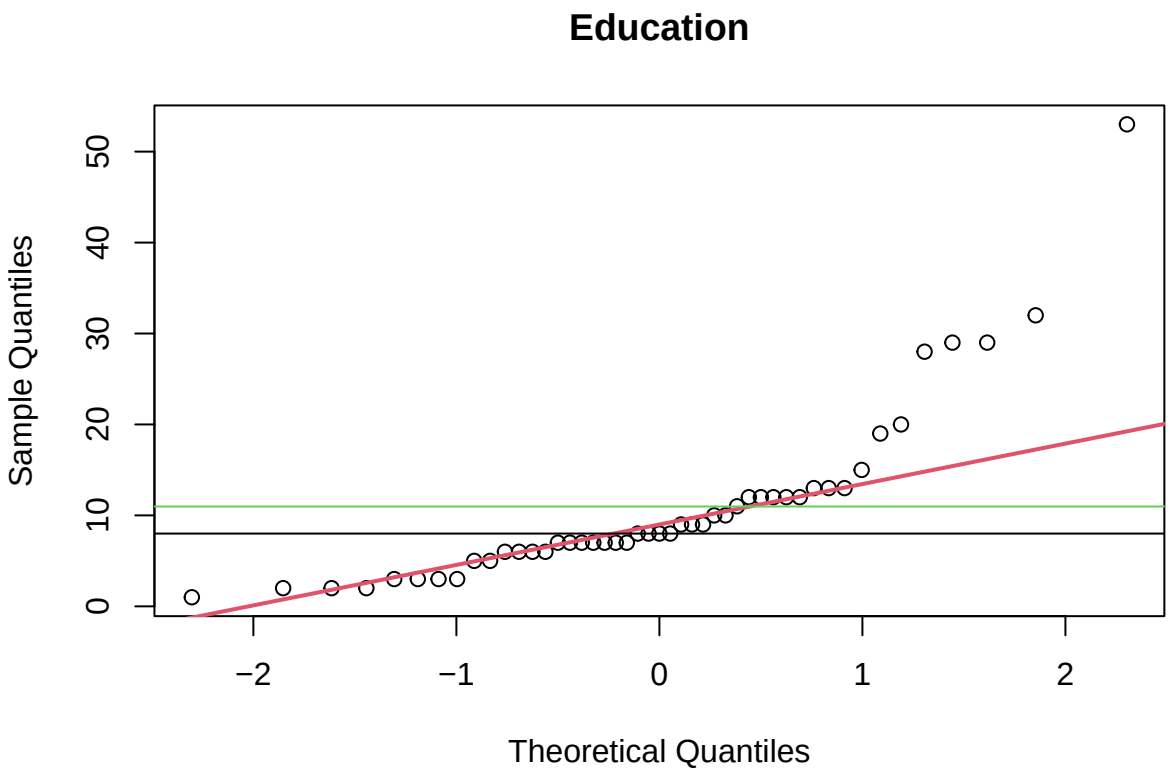

Examination
Density
0.00
0.02
0.04
0
10
20
30
40
Values

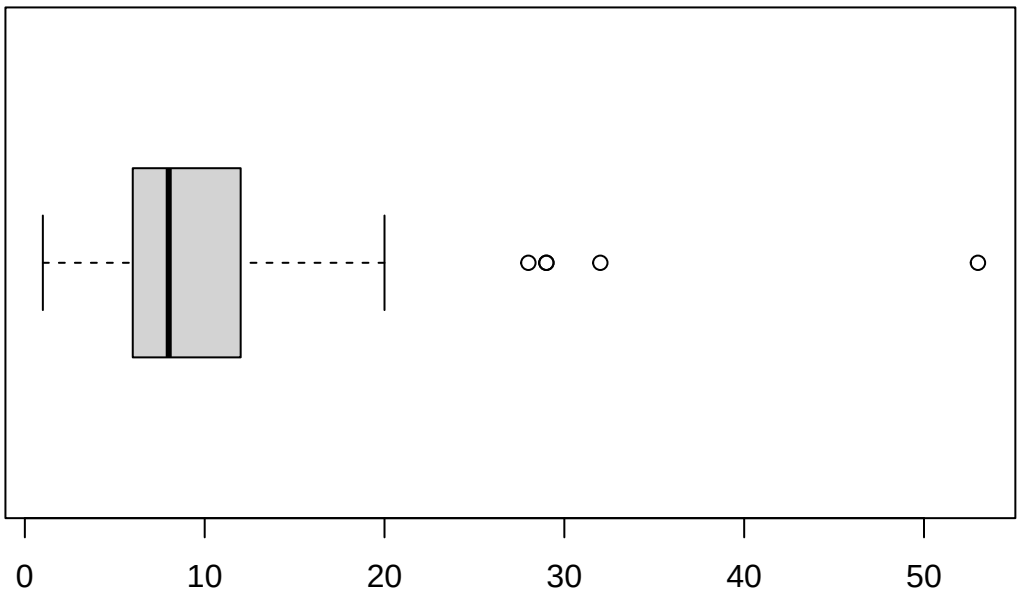

Examination
Sample Quantiles
5
10
15
20
25
30
35
−2
−1
0
1
2
Theoretical Quantiles



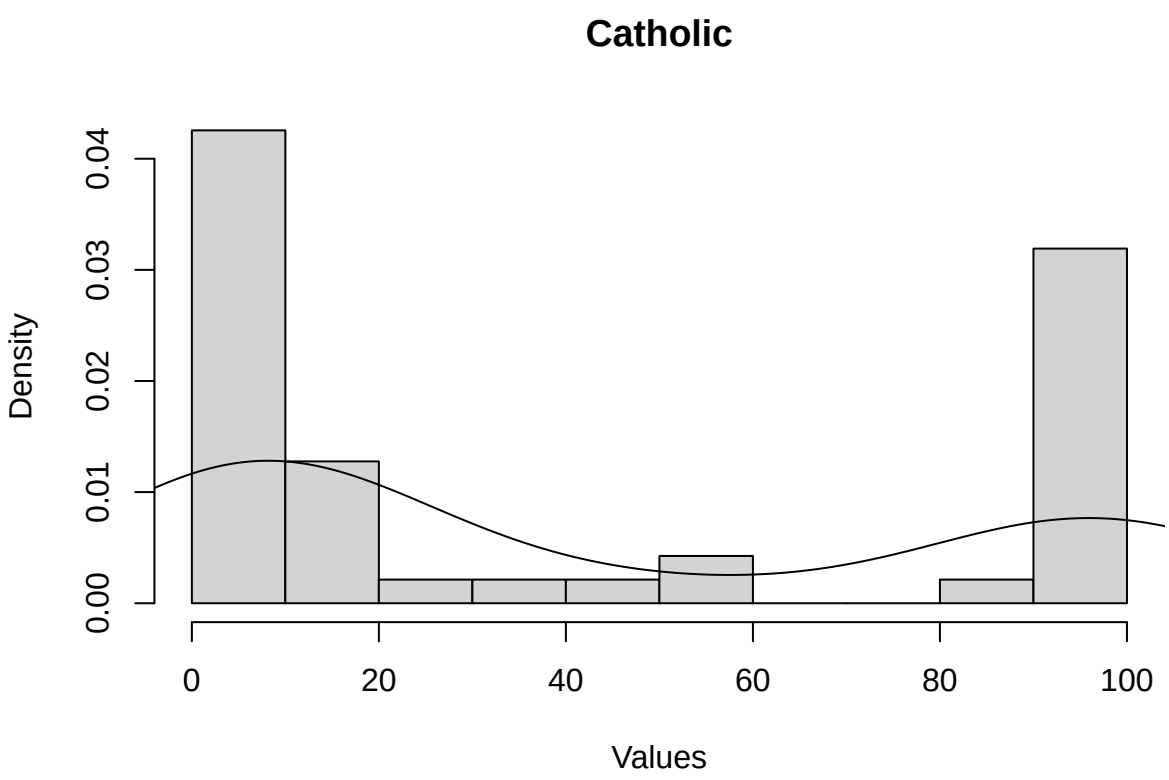

```
## [1] 0.4610349
## [1] 2.988898
## [1] 7.977883
## [1] "Education"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.7482, p-value = 1.312e-07
```

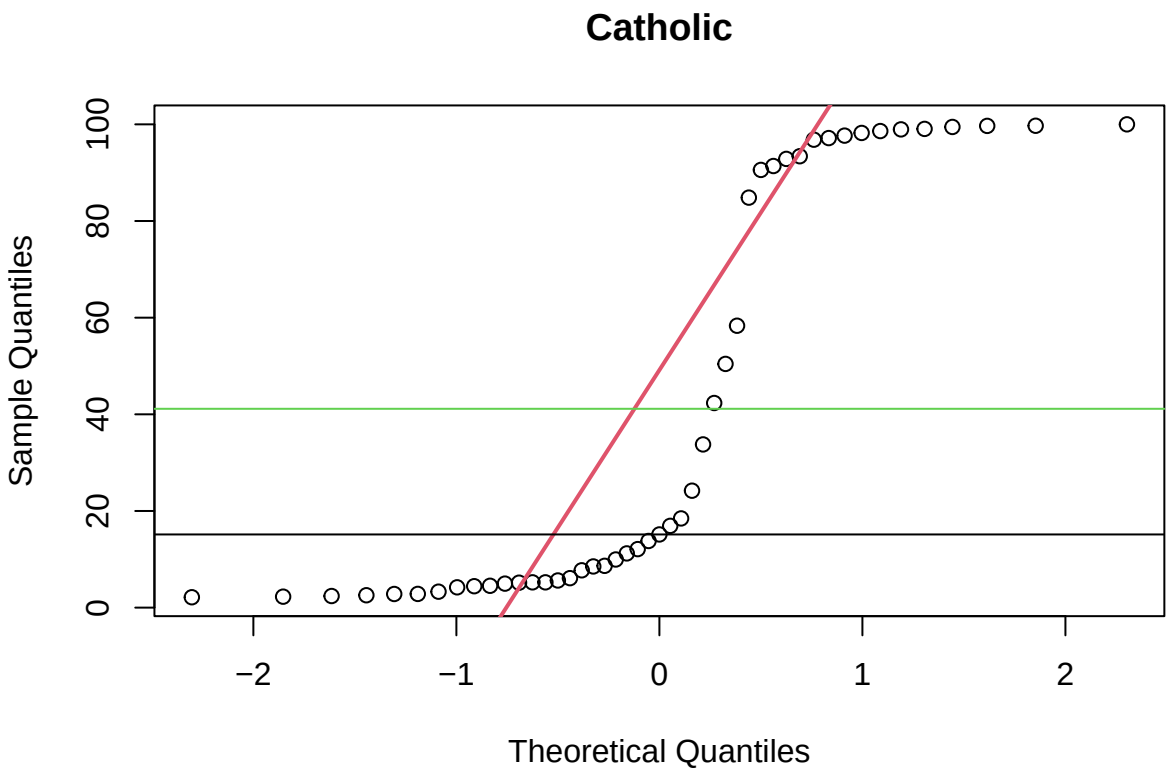

Education
Density
0.00
0.02
0.04
0.06
0
10
20
30
40
50
60
Values

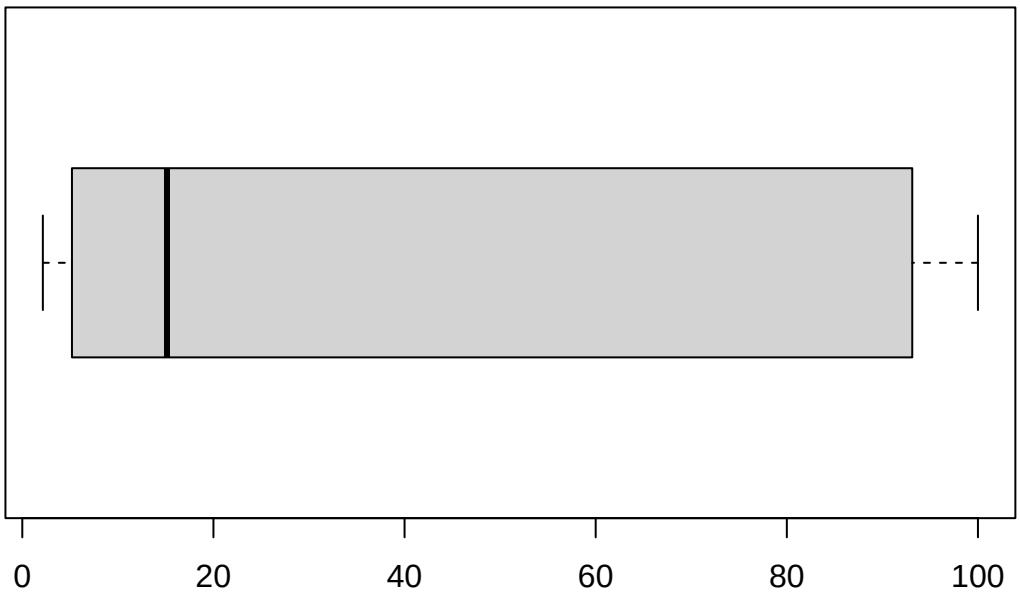

Education
Sample Quantiles
0
10
20
30
40
50
−2
−1
0
1
2
Theoretical Quantiles



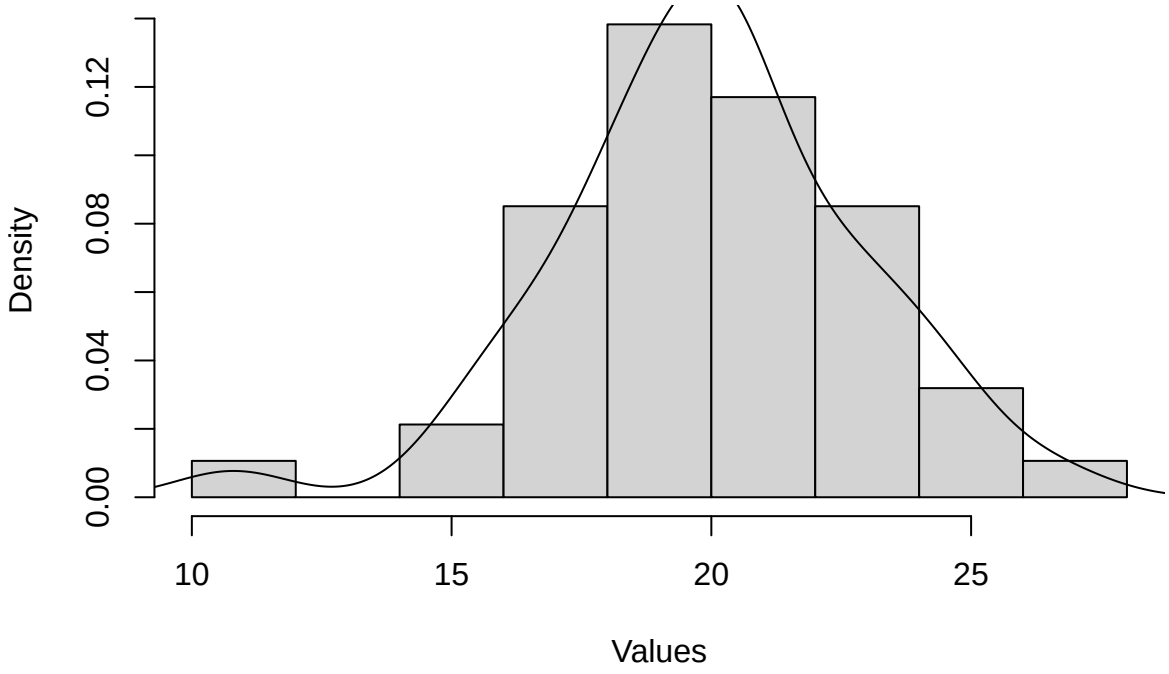

```
## [1] 2.34281
## [1] 9.541434
## [1] 9.615407
## [1] "Catholic"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.7463, p-value = 1.205e-07
```

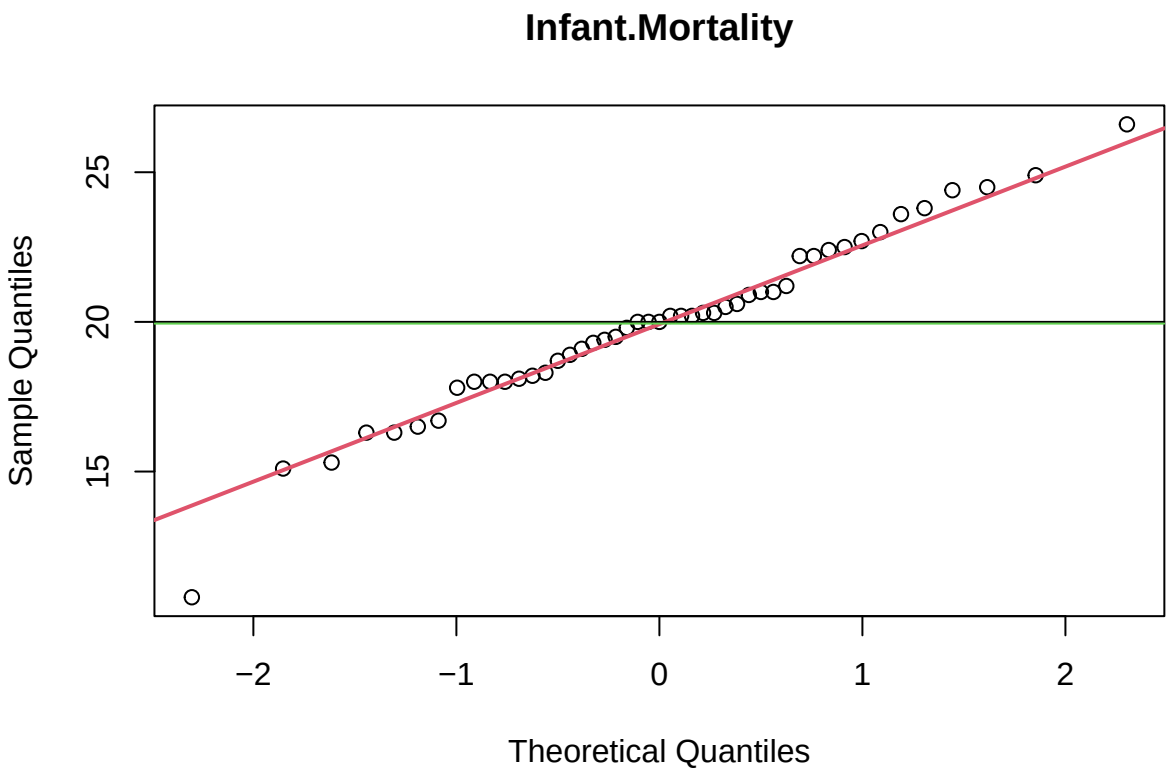

Catholic
Density
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
100
Values

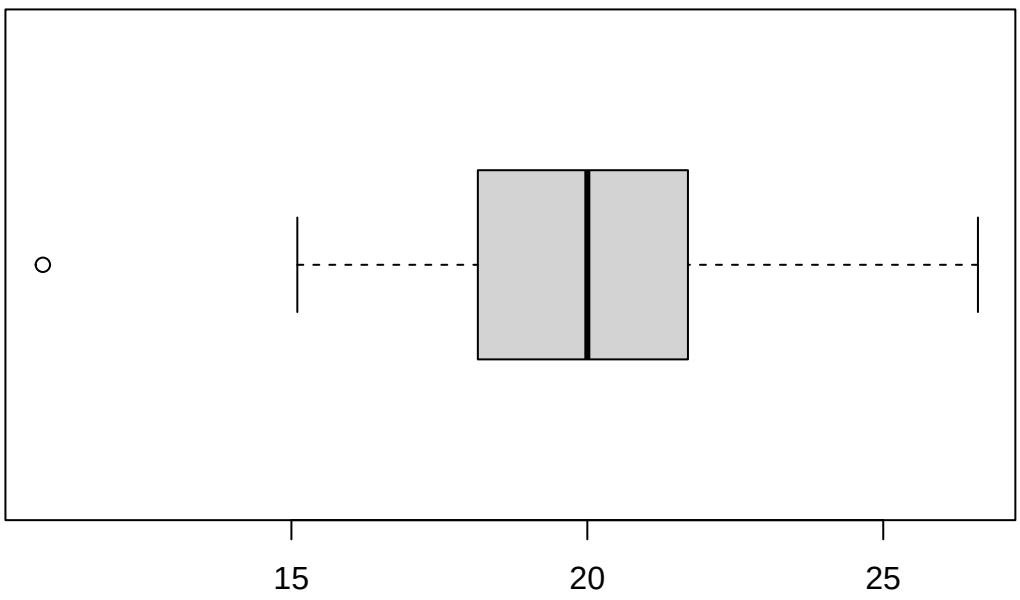

Catholic
Sample Quantiles
0
20
40
60
80
100
−2
−1
0
1
2
Theoretical Quantiles




```
## [1] 0.4946274
## [1] 1.393236
## [1] 41.70485
## [1] "Infant.Mortality"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.97762, p-value = 0.4978
```

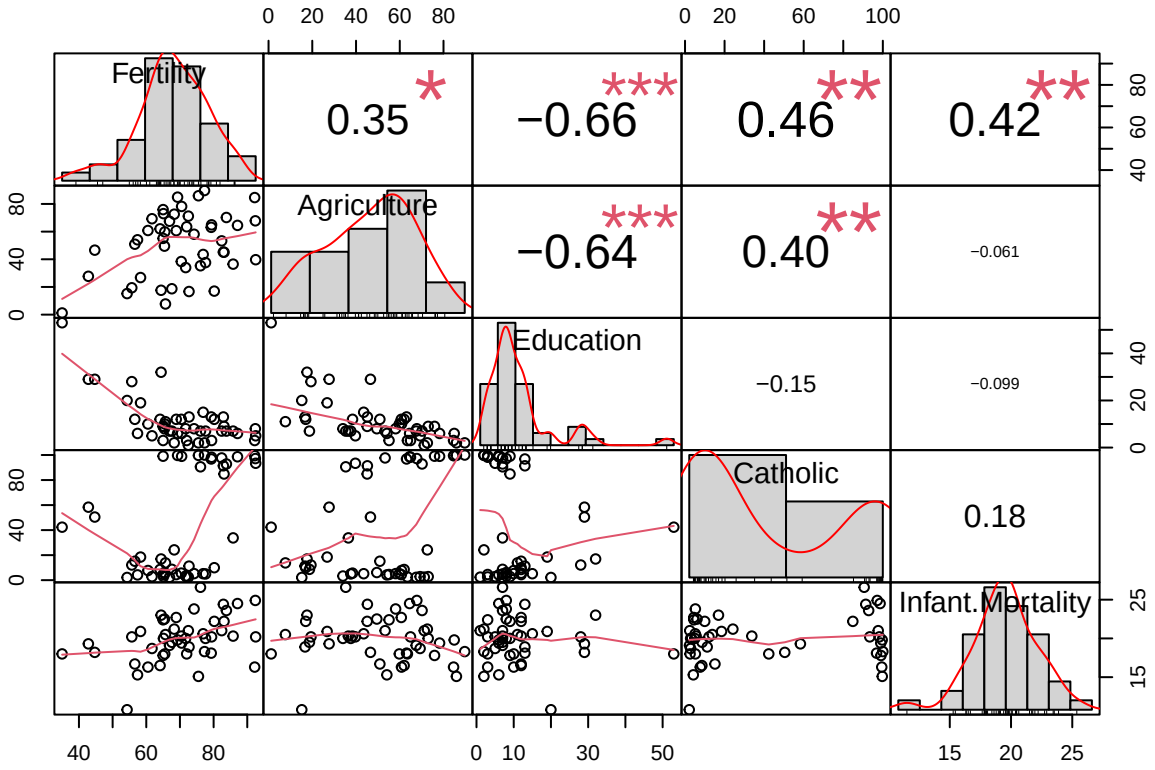

Infant.Mortality
Density
0.00
0.04
0.08
0.12
10
15
20
25
Values


Infant.Mortality
Sample Quantiles
25
20
15
−2
−1
0
1
2
Theoretical Quantiles




```
## [1] -0.3422987
## [1] 3.943302
## [1] 2.912697
```

```
library("PerformanceAnalytics")
my_data <- swiss[, c(1,2,4,5,6)]
chart.Correlation(my_data, histogram=TRUE, pch=19)
```




Die gesammelten Daten beziehen sich auf 47 französischsprachige “Provinzen” um das Jahr 1888.

Alle Variablen außer “Fertility” geben den Anteil an der Bevölkerung an.

Die Untersuchungsvariable “Examination”, die ursprünglich in dem swiss Datenset inkludiert ist, hängt mit Education, Agriculture, Fertility und Catholic stark zusammen, wodurch diese Variable als Regressor für unsere Analyse nicht berücksichtig wurde.

Fertility: Dieser Parameter ist unimodal verteilt. In den meisten Kantonen liegt eine Hohe Fertilität (CSFM) vor. Der Plot lässt auf eine approximative Normalverteilung schließen. Der Kurtosis-Wert von 3,4 bestätigt diese Beobachtung. Der Skwnesswer beträgt -0.47, was eine leichte Linksschiefe impliziert, die auch auf dem Boxplot erkennabar ist. Auch auf dem QQ-Plot ist erkennbar, dass auf der linken Seite ein schwerer Rand existiert. Der Mittelwert ist (beinahe) identisch zum Median, weshalb der Mittelwert als nicht-robustes Lagemaß verwendet werden kann. Das nicht-robuste Streumaß, die Standardabweichung, beträgt 12,49.

Agriculture: Dieser Parameter deutet auf einen hohen Anteil an beschäftigten Männern in der Landwirtschaft hin. Bei diesem Parameter kann von keiner Normalverteilung ausgegangen werden. Der Parameter ist unimodal verteilt. Auf dem QQ-Plot sind keine Ausreißer erkennbar, jedoch weichen die Punkte von den theoretischen Quantilen einer Normalverteilung ab. Als robustes Lagemaß wird der Median verwendet, der 54,10 beträgt.

Education: Der Parameter Education beschreibt die Ausbildung von Männern, die über die Grundschule hinweg geht. Anhand des Boxplots kann man erkennen, dass dieser Parameter unimodal und stark rechtsschief ist. Die kontextuelle Interpretation dieser Verteilung lässt auf einen hohen Anteil an Männern mit geringer Bildung schließen. Der QQ-Plot zeigt schwere Ränder auf der rechten Seite. Für diesen Parameter empfiehlt sich ebenfalls der Median als robustes Lagemaß und beträgt 8,00.

Catholic: Dieser Parameter gibt die prozentuelle Zahl an Katholiken (C) in den Kantonen an, wobei die Anzahl an Protestanten 1-C beträgt. Dieser Parameter ist bimodal für Werte zwischen 0-10 und 90-100, was darauf hindeutet, dass Kantone entweder katholisch oder prostestantisch waren. Somit entfallen Median

und arithmetisches Mittel als sinnvolle Lagemaße, was auch durch die Länge des Boxplots verdeutlicht wird. Die Interquartilsdistanz deckt beinahe den vollständigen Wertebereich ab.

Infant Mortality: Dieser Parameter gibt die Säuglingssterblichkeit an und wird definiert als "Lebendgeborene, die weniger als 1 Jahr leben". Dieser Parameter scheint approximativ normalverteilt zu sein, was im realen Kontext damit begründet wird, dass sich Säuglingssterblichkeit in Kantonen, die medizinisch und entwicklungstechnisch nicht übermässig verschieden sind, um einen Mittelwert einpendelt. Der Kurtosis-Wert von 3,94 impliziert eine steilgipflige Normalverteilung. Auf dem QQ-Plot ist zu erkennen, dass (fast) alle Werte nahe der theoretischen Quantil-Linie lokalisiert sind. Am linken Rand ist ein Ausreißer erkennbar. Dennoch nehmen wir hier eine Normalverteilung an und analysieren mittels nicht-robusten Mittelwert von 19,94 und einer Standardabweichung von 2,91.

**Korrelation der Parameter**

Korrelationsmatrix: Anhand der Korrelationsmatrix kann die Korrelation der einzelnen Parameter miteinander verglichen werden. Dabei korrelieren Fertility~Education mit -0.66 und Agriculture~Education mit -0.64 negativ zueinander. Diese Entwicklung passt sozio-ökonomisch zu modernen Entwicklungen, wobei Länder mit steigender Bildung einen Geburtenrückgang zeigen. Auch der Rückgang von Männern in der Landwirtschaft geht mit steigender Bildung zurück, wobei dies auch mit steigenden technischen Innovationen zusammenhängen könnte. Aus dem Parameter "Catholic" lassen sich aufgrund der Bimodalität keine sinnvolle Korrelation mit anderen Parametern ableiten.

**Aufgabe 2**

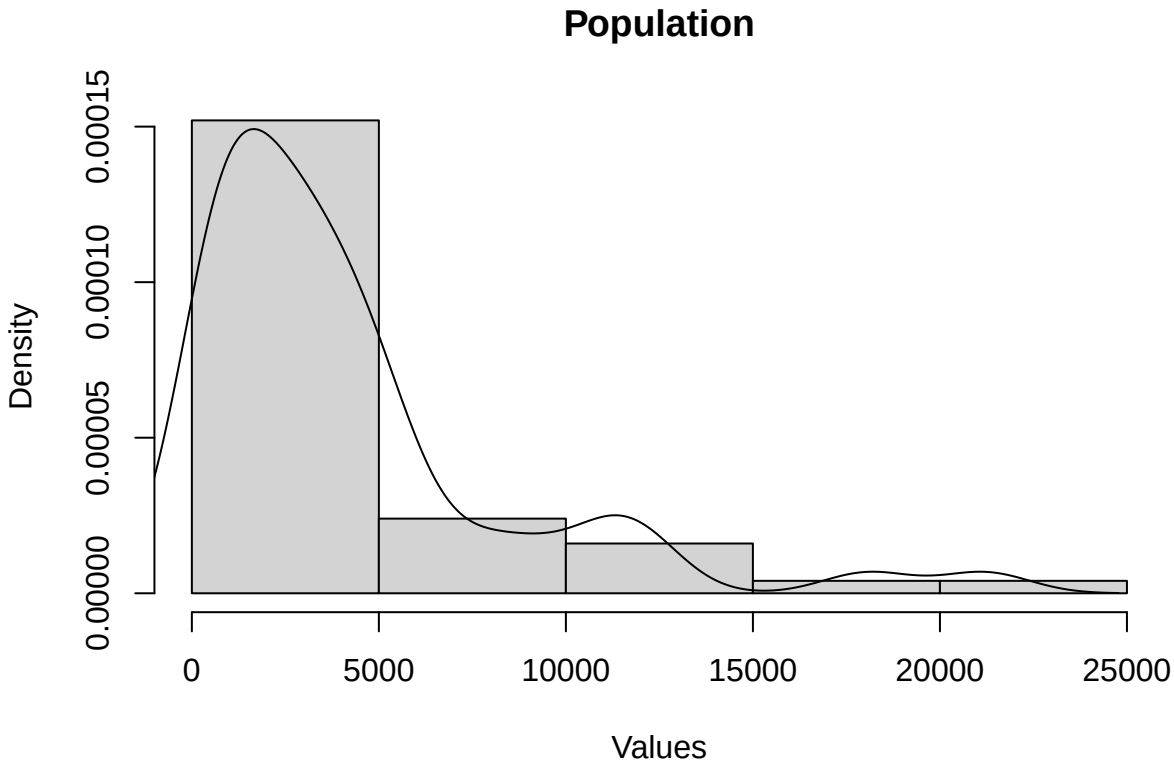
Explorieren und visualisieren Sie die Variablen “Population”, “Income”, “Illiteracy”, “Life.Exp”, “Murder”, “HS Grade” und “Frost” aus dem R Datensatz state.x77. Betrachten Sie vorerst jede Variable als separate Stichprobe für eindimensionale Exploration (ziehen Sie die Bedeutung der Variablen im Sachkontext in Betracht). Fur jede Variable: • wählen Sie sinnvolle Schätzer für Lokation, Variation, Schiefe and Gewicht in den Rändern. • Geben Sie dem Nutzer die Möglichkeit zwischen unterschiedlichen graphischen Darstellungen zu wechseln. Erklären Sie die Zusammenhänge und Eiegenschaften der Daten, die sich aus diesen Visualisierungen erkennen lassen. – Sind die Daten symmetrisch/schief? – Haben die Daten schwere Ränder? – Bieten Sie robuste und nichtrobuste Lagemaße und Skalenmaße im Vergleich oder zur Auswahl an. – Sind die Daten (approximativ) normalverteilt? Was lässt sich uber die Zusammenhänge zwischen den Variablen aussagen? (Tipp: scatterplot matrix.)

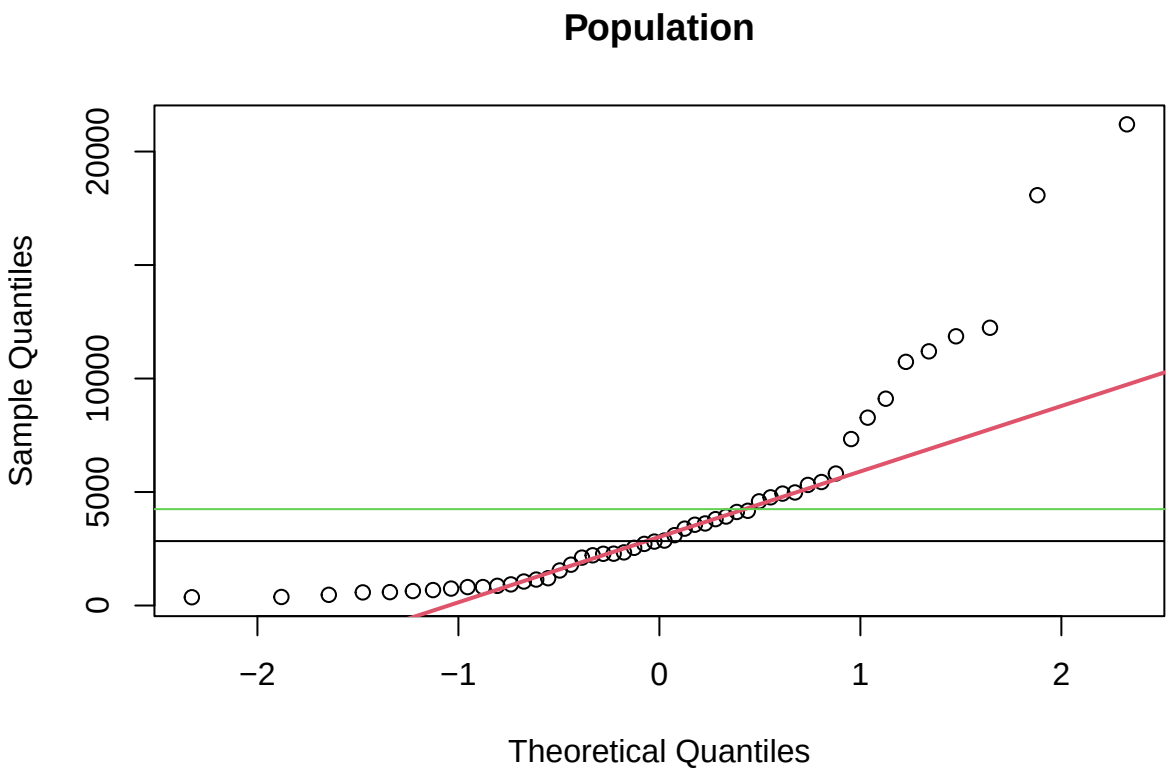
```
data = data.frame(state.x77)
summary(data)
```

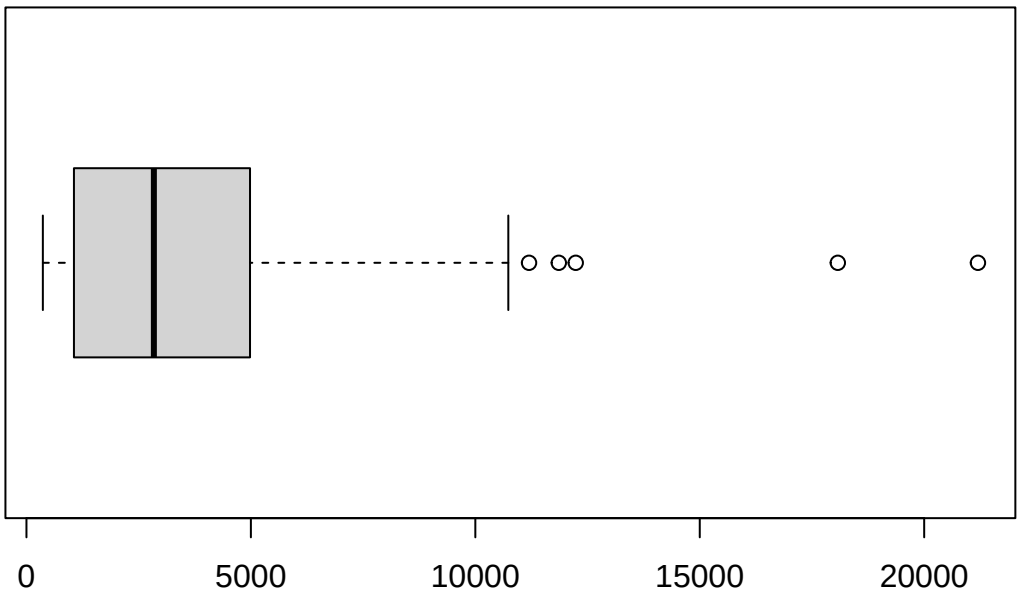
```
##    Population        Income       Illiteracy       Life.Exp
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
##  Median : 2838   Median :4519   Median :0.950   Median :70.67
##  Mean   : 4246   Mean   :4436   Mean   :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.: 81163
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
```

```
df <- state.x77

i= 1
for (k in data){
  mycols <- colnames(df)

  i<-i+1
}

for (k in (1:7)) {

  print(mycols[k])
  print(shapiro.test(data[,k]))
  hist(data[,k],main = paste(mycols[k]),freq = F, xlab = "Values"); lines(density(data[,k]))
  qqnorm(data[,k], main = paste(mycols[k]))
  qqline(data[,k],col=2,lwd=2)
  abline(h=median(data[,k]))
  abline(h=mean(data[,k]),col=3)
  boxplot(data[,k], horizontal = T, main = paste(mycols[k]))
   print(skewness(data[,k]))
  print(kurtosis(data[,k]))
  print(sd(data[,k]))
}
```

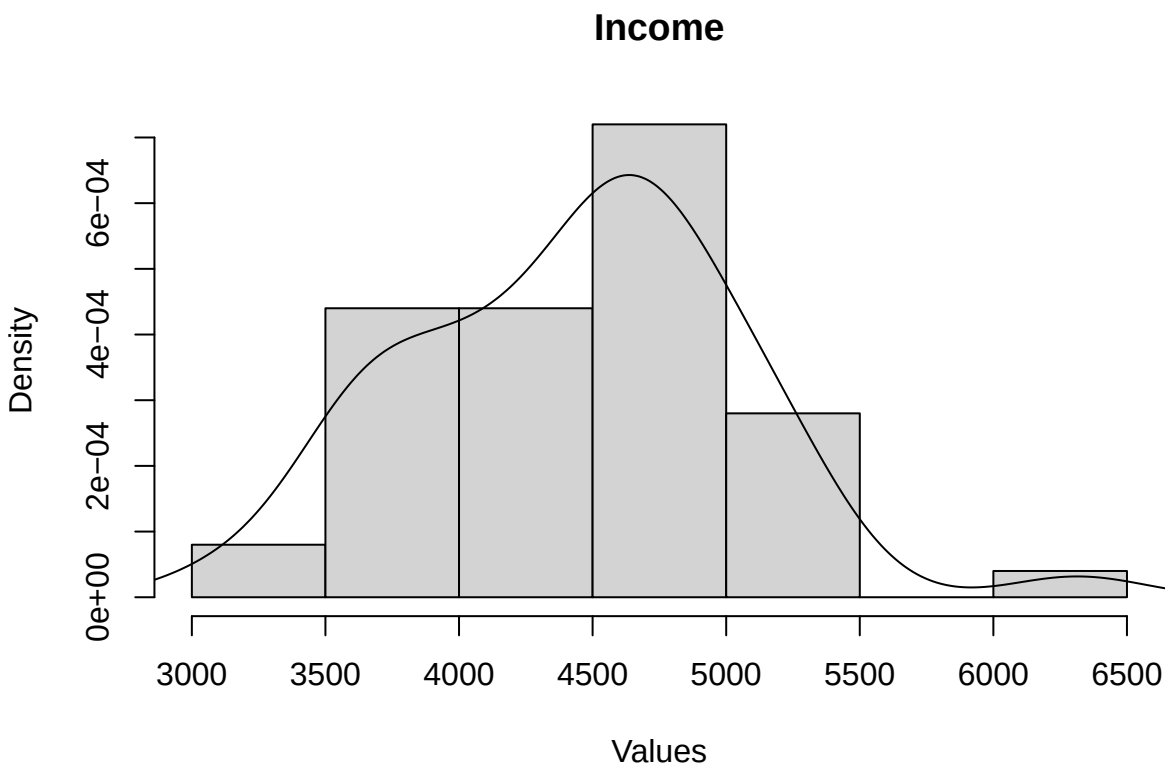
```
## [1] "Population"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.76999, p-value = 1.906e-07
```

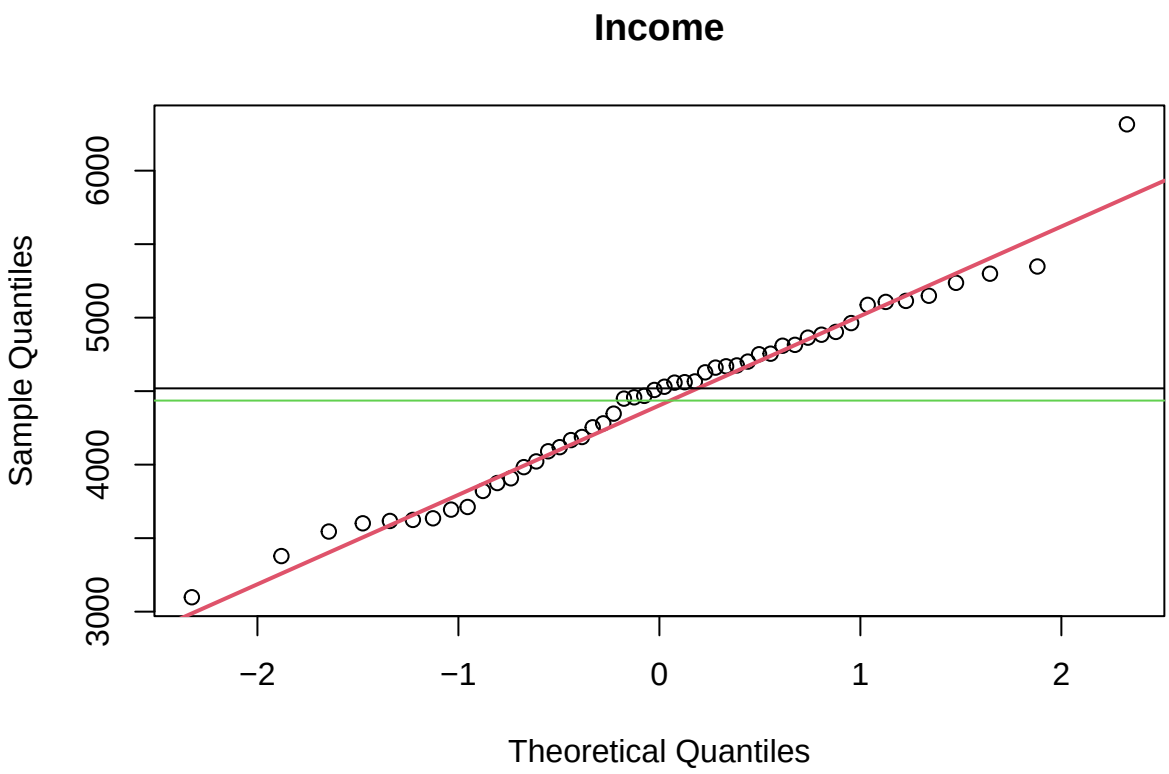


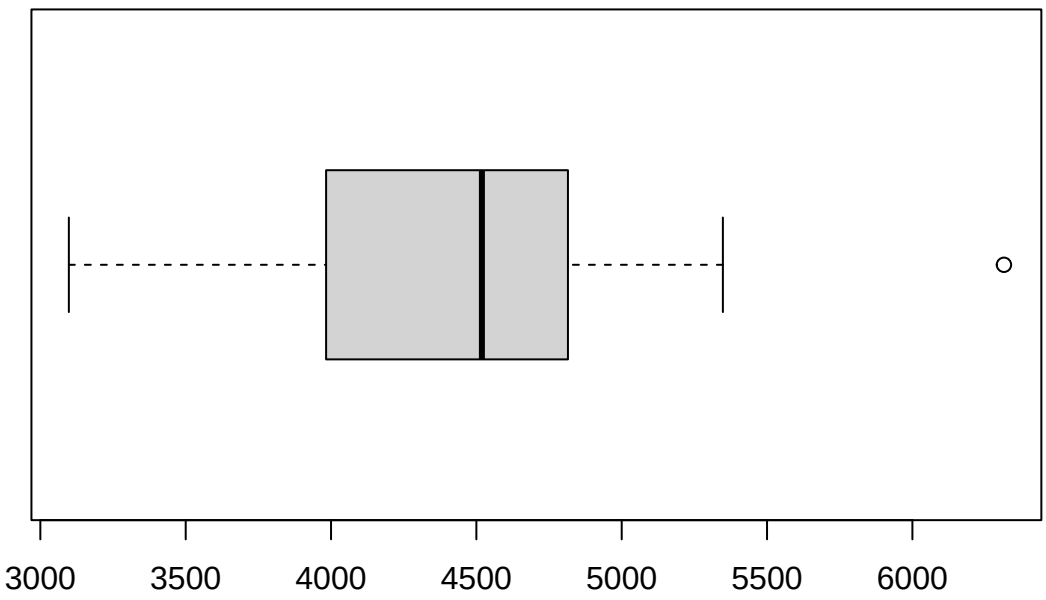

Population
Sample Quantiles
0
5000
10000
20000
−2
−1
0
1
2
Theoretical Quantiles



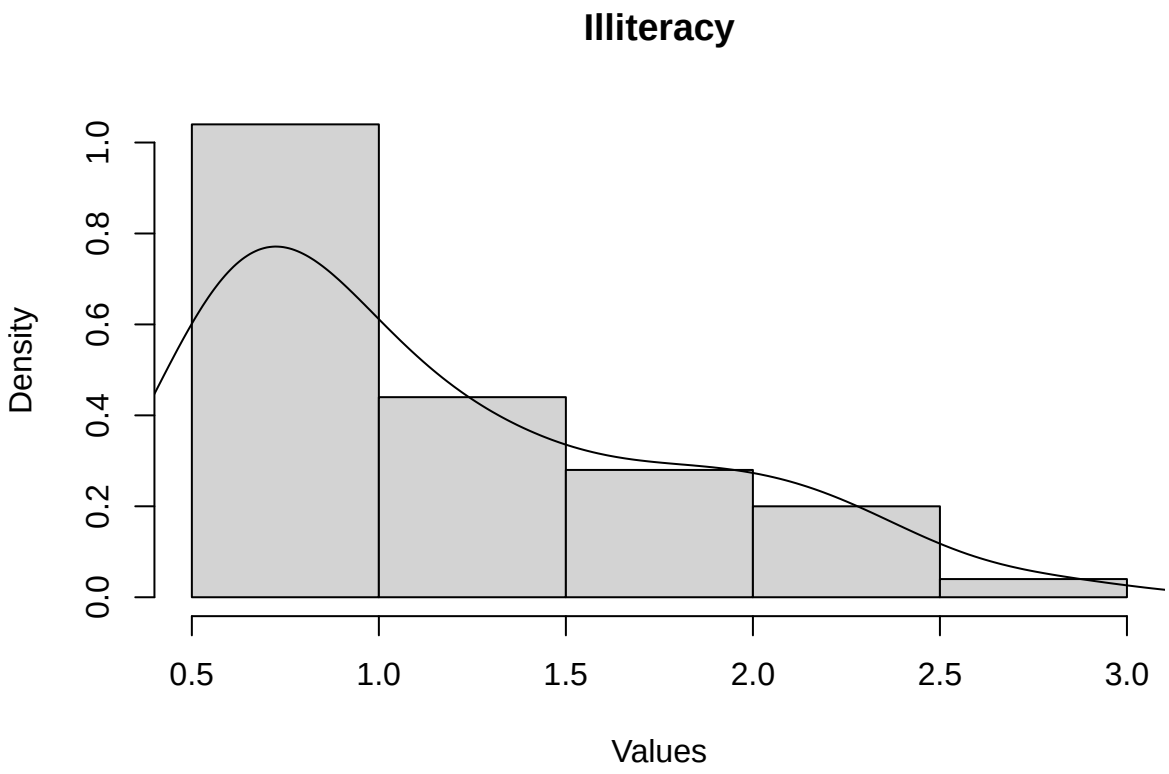

```
## [1] 1.981395
## [1] 4.030555
## [1] 4464.491
## [1] "Income"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.9769, p-value = 0.43
```

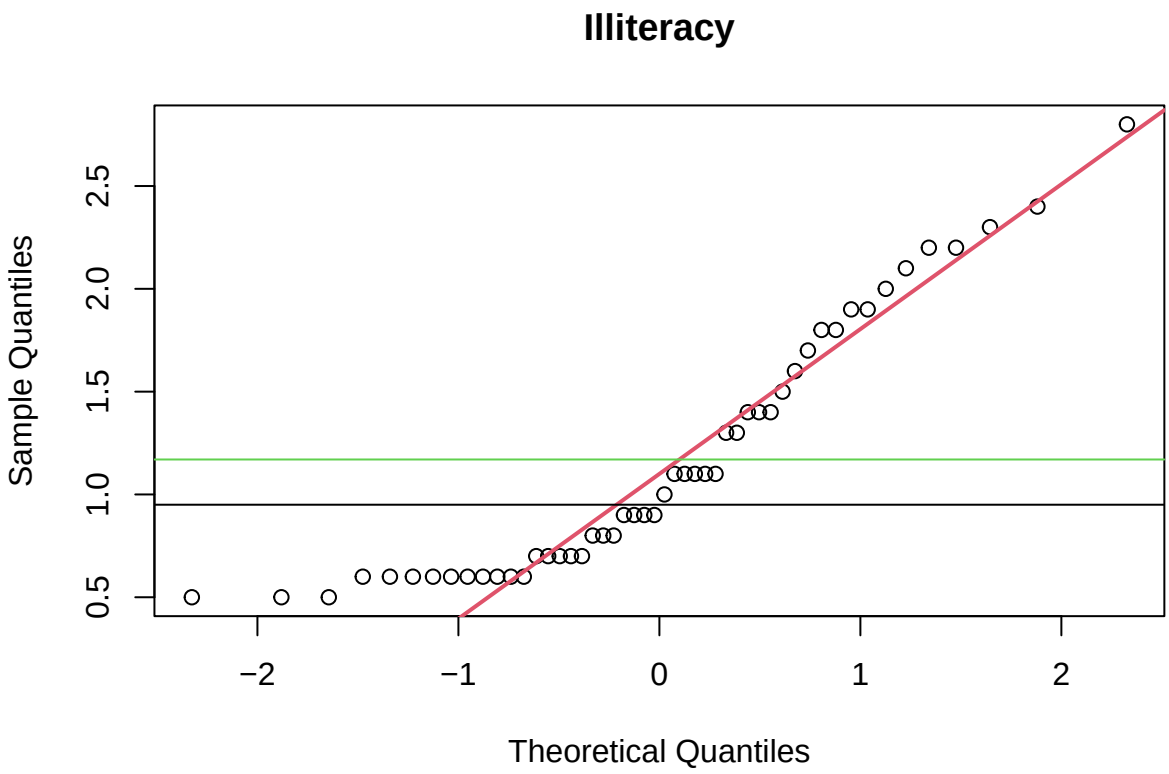

Income
Density
0e+00
2e−04
4e−04
6e−04
3000
3500
4000
4500
5000
5500
6000
6500
Values

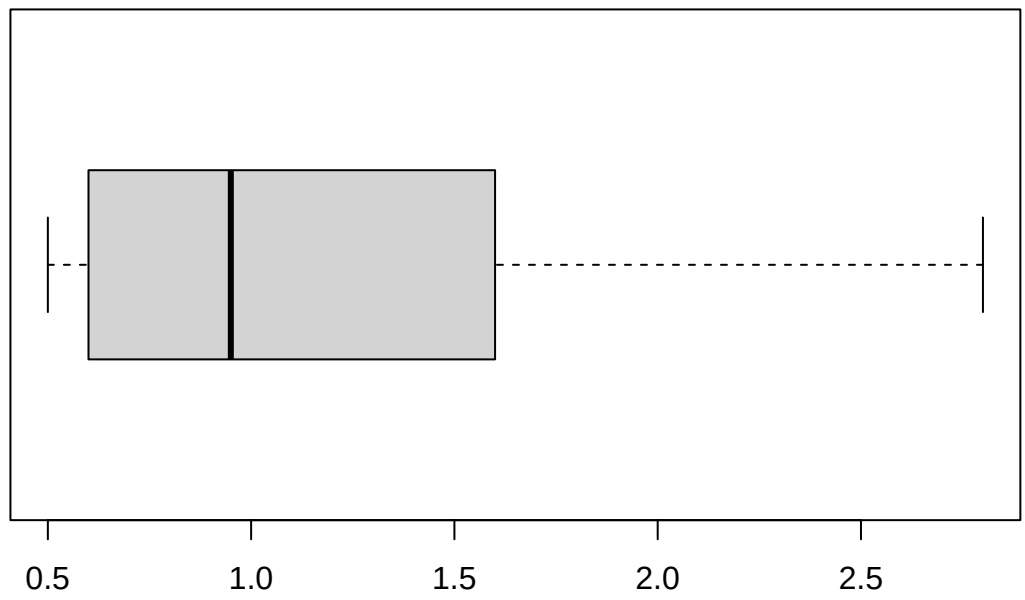

Income
Sample Quantiles
3000
4000
5000
6000
−2
−1
0
1
2
Theoretical Quantiles



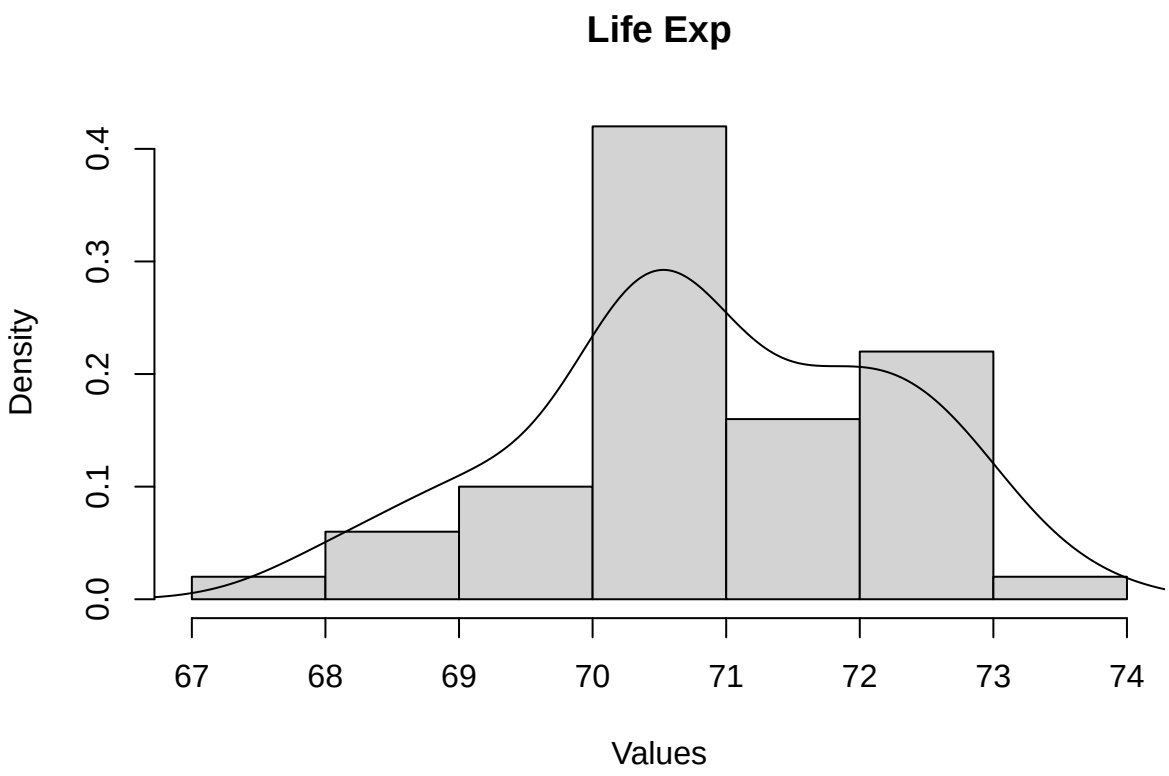

```
## [1] 0.2109882
## [1] 0.3783528
## [1] 614.4699
## [1] "Illiteracy"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.88315, p-value = 0.0001396
```

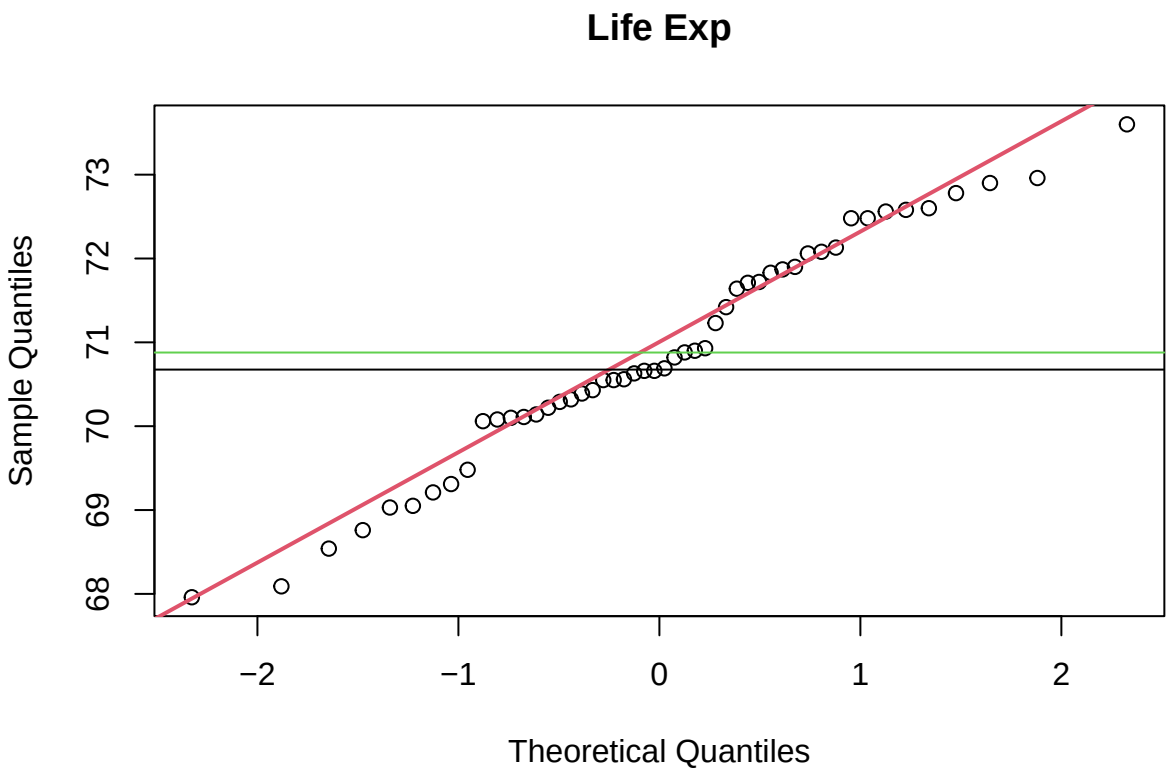

Illiteracy
Density
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
2.5
3.0
Values

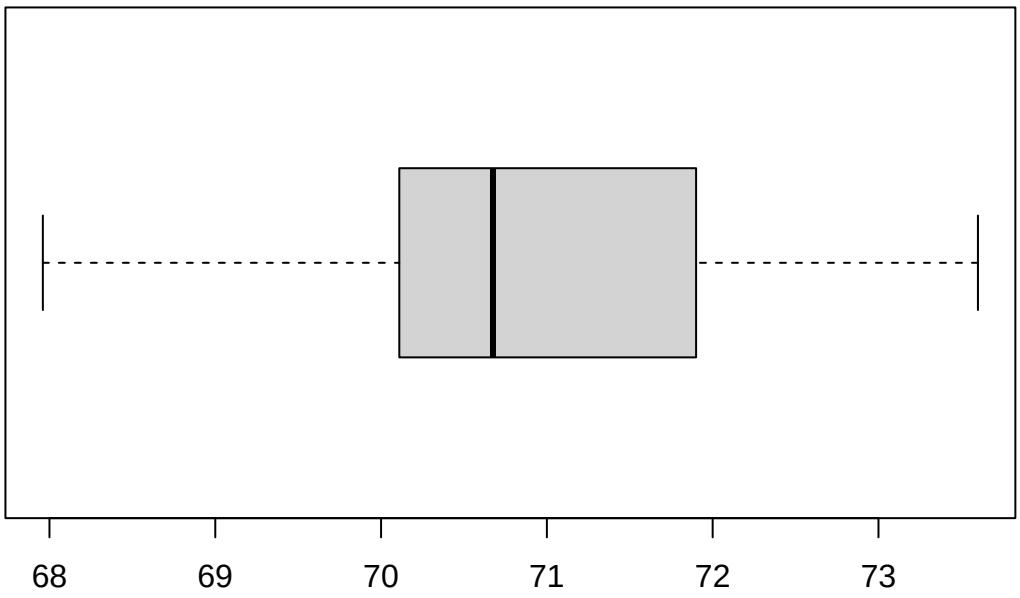

Illiteracy
Sample Quantiles
0.5
1.0
1.5
2.0
2.5
−2
−1
0
1
2
Theoretical Quantiles



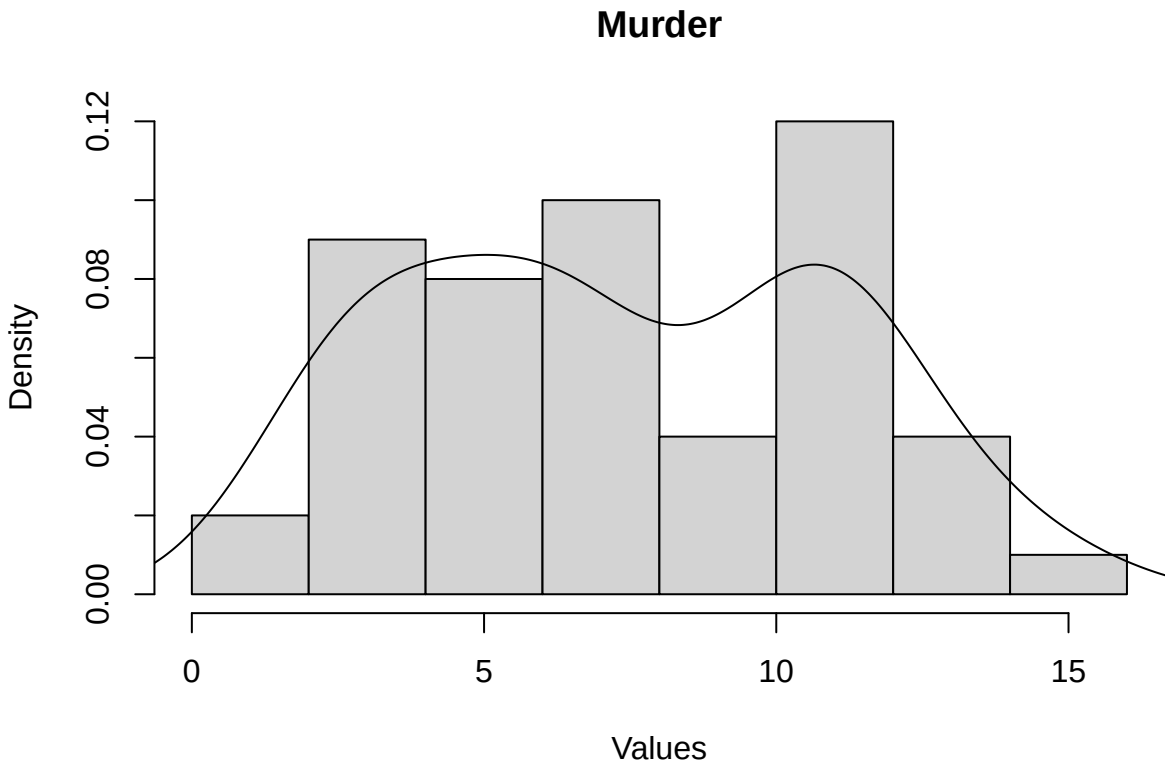

```
## [1] 0.8437669
## [1] -0.3671622
## [1] 0.6095331
## [1] "Life Exp"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.97724, p-value = 0.4423
```

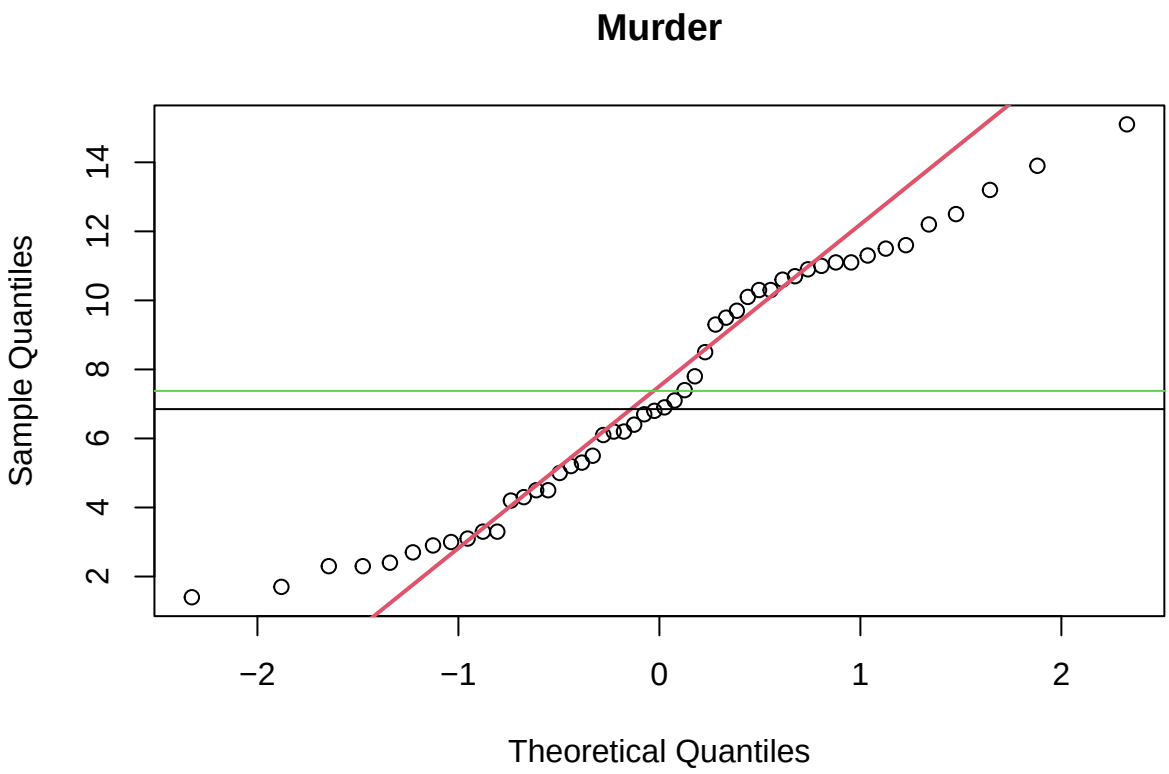

Life Exp
Density
0.0
0.1
0.2
0.3
0.4
67
68
69
70
71
72
73
74
Values

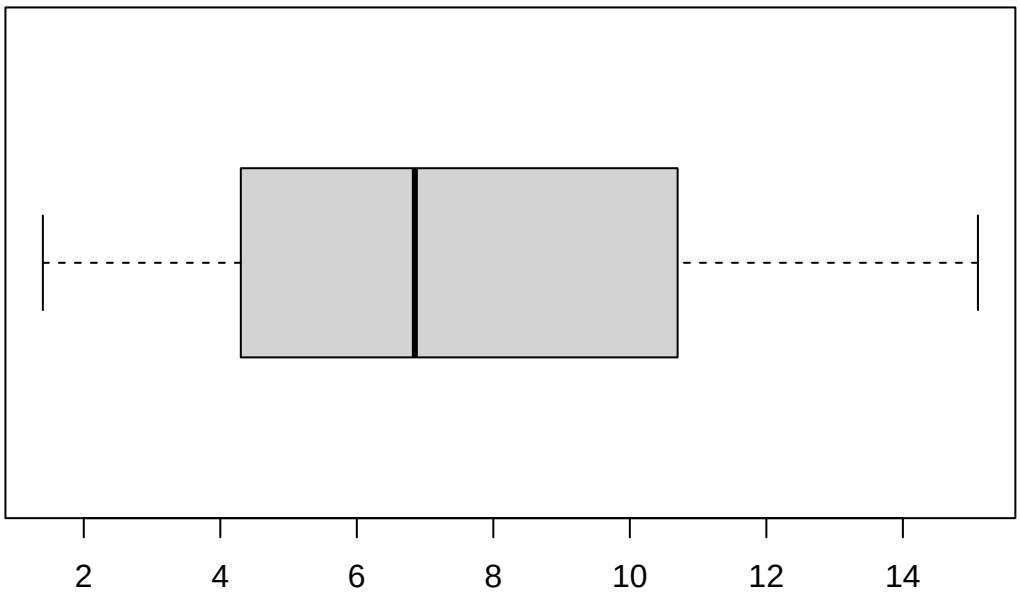

Life Exp
Sample Quantiles
68
69
70
71
72
73
−2
−1
0
1
2
Theoretical Quantiles



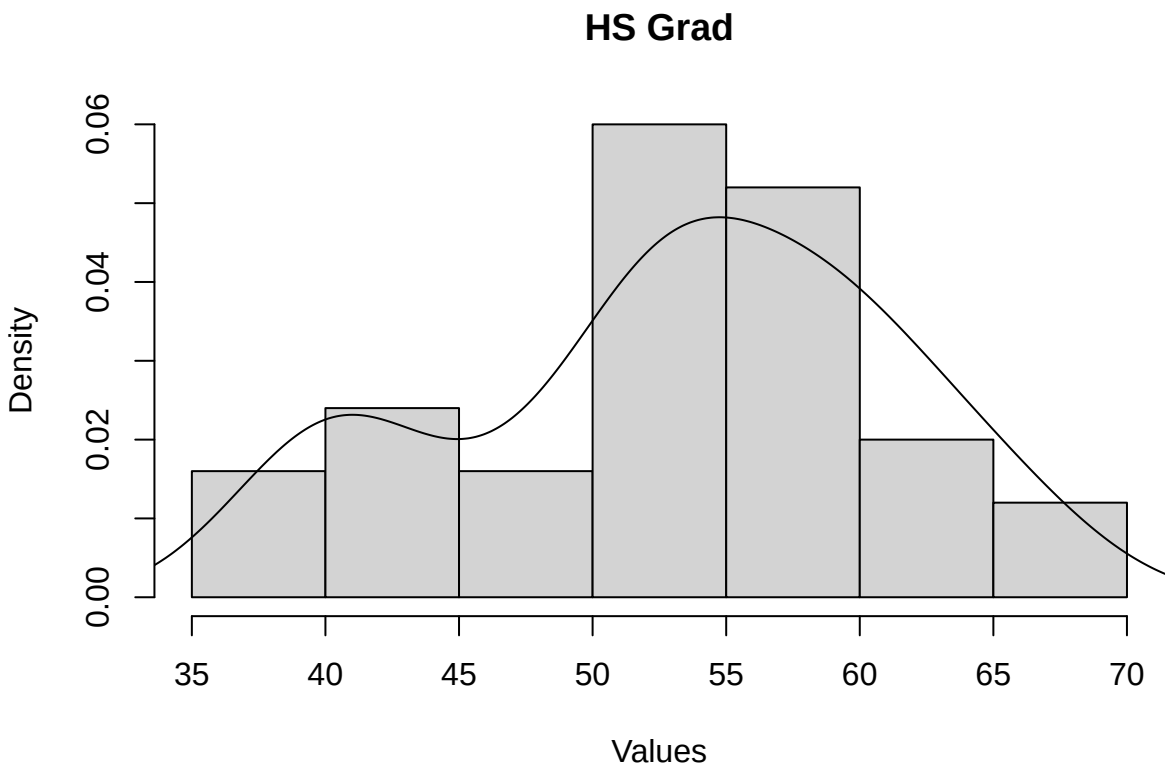

```
## [1] -0.1582224
## [1] -0.5729621
## [1] 1.342394
## [1] "Murder"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.95347, p-value = 0.04745
```

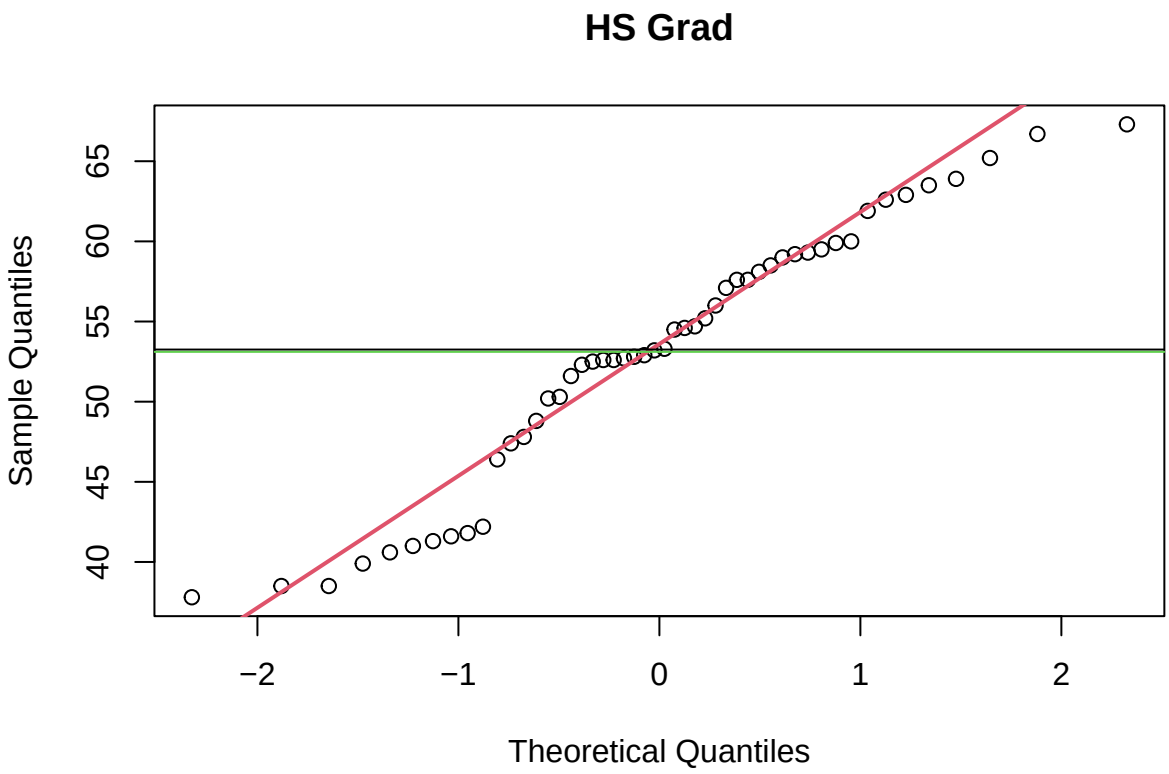

Murder
Density
0.00
0.04
0.08
0.12
0
5
10
15
Values

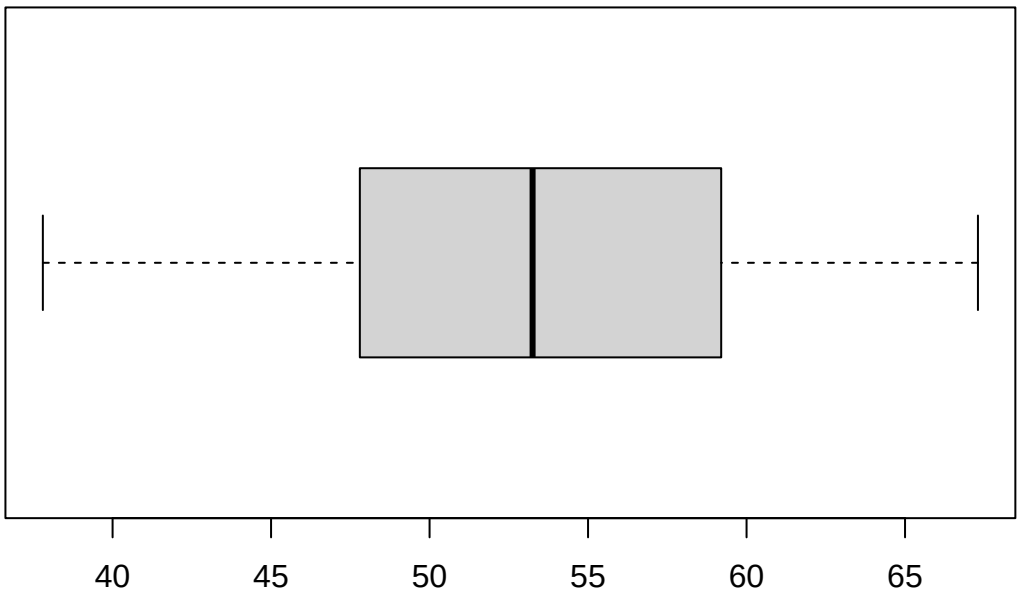

Murder
Sample Quantiles
2
4
6
8
10
12
14
−2
−1
0
1
2
Theoretical Quantiles



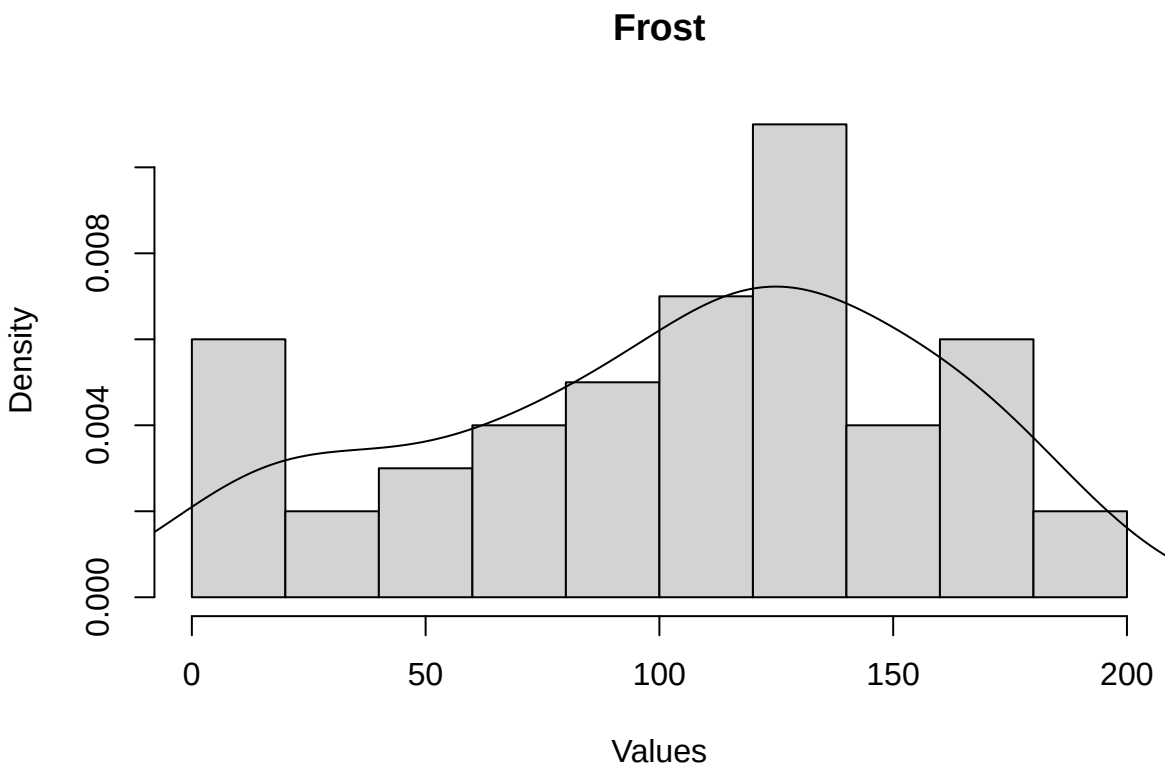

```
## [1] 0.1333186
## [1] -1.137441
## [1] 3.69154
## [1] "HS Grad"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.9531, p-value = 0.04582
```

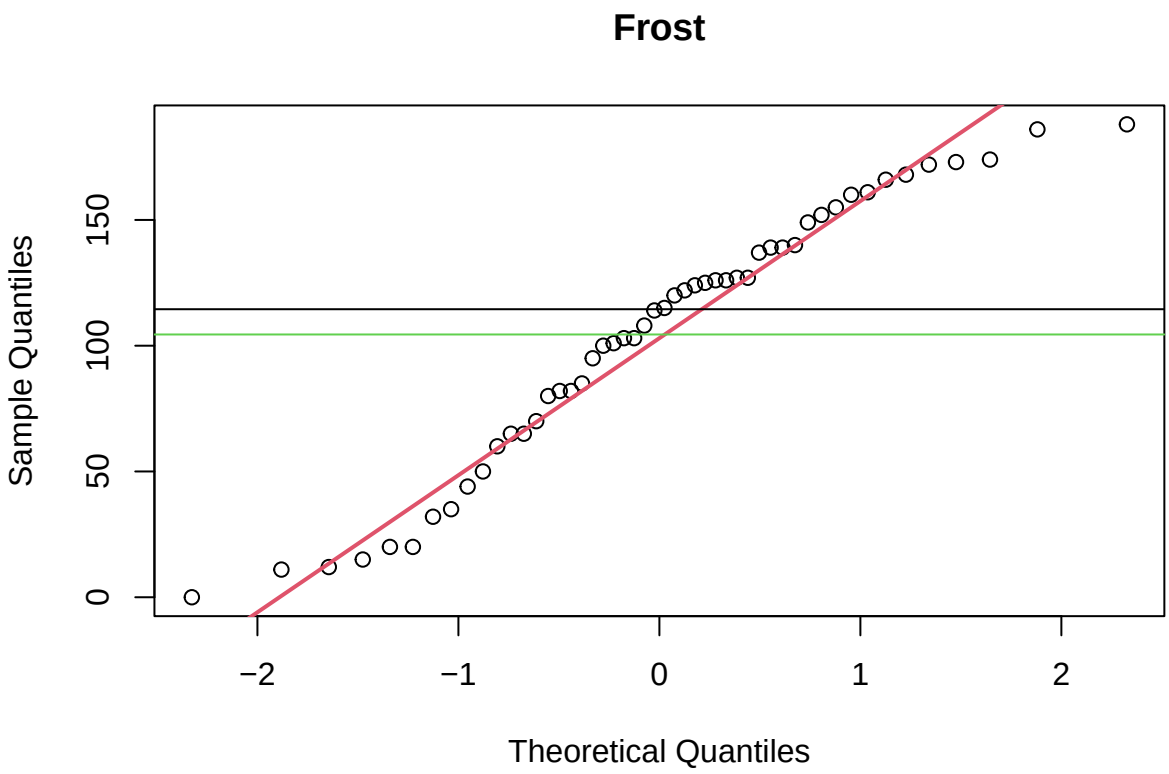

HS Grad
Density
0.00
0.02
0.04
0.06
35
40
45
50
55
60
65
70
Values

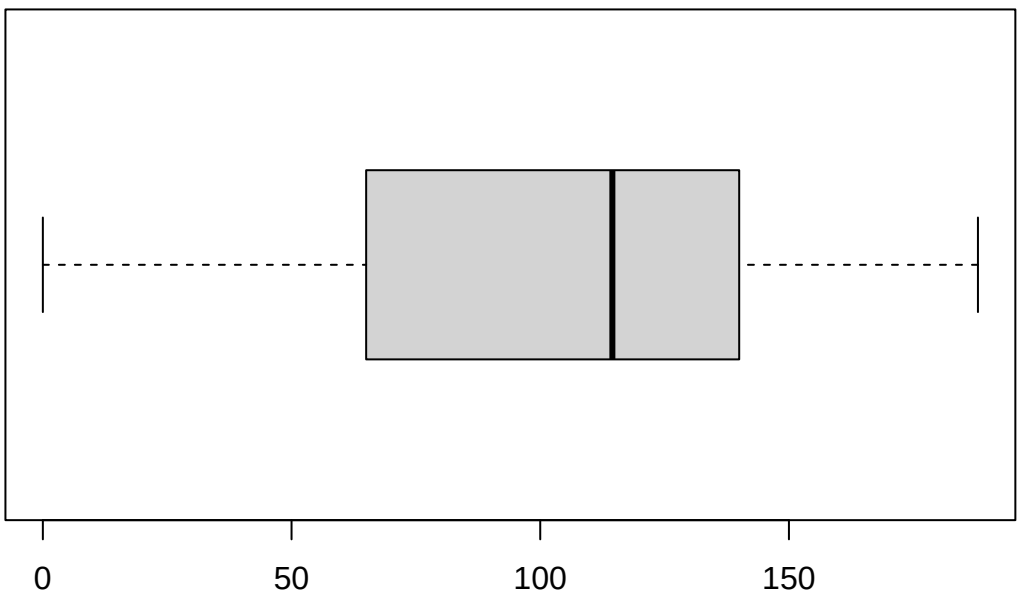

HS Grad
Sample Quantiles
40
45
50
55
60
65
−2
−1
0
1
2
Theoretical Quantiles




```
## [1] -0.3290666
## [1] -0.7904271
## [1] 8.076998
## [1] "Frost"
##
##  Shapiro-Wilk normality test
##
## data:  data[, k]
## W = 0.95456, p-value = 0.05267
```

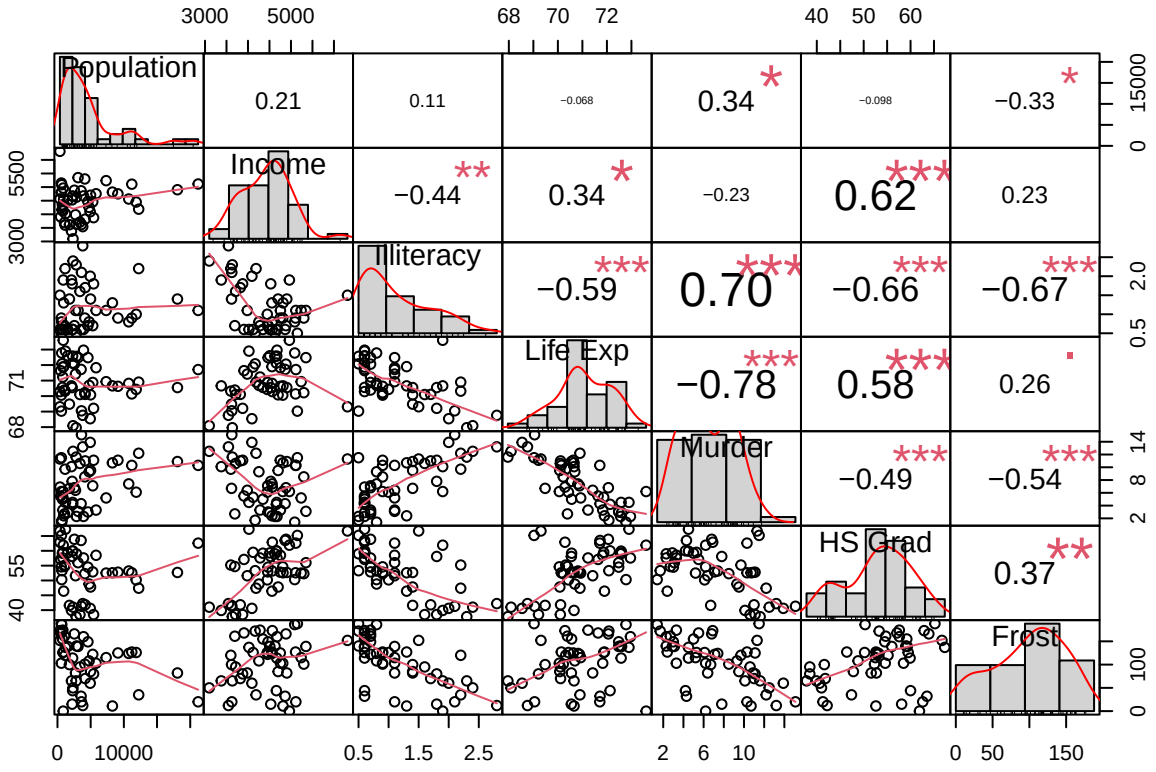

Frost
Density
0.000
0.004
0.008
0
50
100
150
200
Values


Frost
Sample Quantiles
0
50
100
150
−2
−1
0
1
2
Theoretical Quantiles




```
## [1] -0.3776493
## [1] -0.859056
## [1] 51.98085
```

```
library("PerformanceAnalytics")
my_data <- state.x77[, c(1,2,3,4,5,6,7)]
chart.Correlation(my_data, histogram=TRUE, pch=19)
```




Der Datensatz “state.x77” besitzt 7 unterschiedliche Parametern zu den 50 Bundesstaaten der Vereinigten Staaten von Amerika, nämlich “Population”, “Income”, “Illiteracy”, “Life.Exp”, “Murder”, “HS Grade” und “Frost”.

Population: Der Parameter ist die geschätzte Population der 50 US-Bundesstaaten im Jahr 1975. Die Verteilung im Histogramm ist unimodal, zeigt eine deutliche rechtsschiefe mit einem schweren Rand auf der rechten und einem leichter Rand auf der linken Seite. Auch im QQ-Plot ist diese Beobachtung ersichtlich. Eine Normalverteilung ist damit auszuschließen. Dieser wäre auch im Hinblick auf die flächenmässige Diskrepanz der Bundesstaaten nicht zu erwarten. Für eine derartige Verteilung eignet sich der nicht-robuste Mittelwert nicht. Hierbei ist der robuste Median zu wählen, mit dem Wert 2838. Das Boxplot-Diagramm ist hierbei eine sinvolle Illustration, die zeigt, dass in 5 US-Staaten die Bevölkerung stark über den anderen Staaten liegt.

Income: Der Income Parameter beschreibt das Pro-Kopf Einkommen. Dieser Wert ist approximativ normalverteilt. Die Punkte liegen dicht um die Normalverteilungsgerade des QQ-Plots, nur ein Ausreißer auf der rechten Seite ist erkennbar. Der Median und das arithmetische Mittel liegen nah beieinander, weshalb das arithmetische Mittel von 4436 mit einer Standardabweichung von 614.4699 sinnvolle Lage- und Streumaße sind.

Illiteracy: Dieser Paramater gibt die prozentuelle Anzahl an Analphabeten an. Diese Verteilung ist rechtschief mit einem deutlichen schweren Rand auf der linken Seite. Die positive Skewness-Wert von 0.8437669 bestätigt die Rechtsschiefe. Das robuste Lagemaß Median scheint mit 0,95 hier der sinnvollste Lageschätzer, da das arithmetische Mittel von diesem um 20 Prozentpunkte abweicht.

Life.Exp: Dieser Parameter beschreibt die Lebenserwartung in Jahren. Es ist eine leichte Bimodalität zu erkennen. Es sind keine schweren Ränder ersichtlich. Als Lageschätzer wird aufgrund der Bimodalität der Median herangezogen, welcher 70.67 beträgt. Das arithmetische Mittel liegt zwar knapp mit 70.88 daneben, allerding ist dieses bei einer gegebenen Bimodalität zu verwerfen. Der Boxplot kann hierbei als sinnvolle Visualisierung der Daten-Streuung angesehen werden.

Murder: Murder gibt die Mord- und Totschlagrate pro 100.000 Einwohner an. Hier liegt ebenfalls eine bimodale Verteilung vor. Analog zu Life Exp. wird der Median mit 6.850 als zentraler Lageschätzer herangezogen.

HS Grade: Dieser Parameter gibt die prozentuelle Anzahl der High-School Absolventen an. Die Daten haben auf dem Histogramm eine bimodale Form, wobei der erste Peak nicht sehr stark ausgepägt ist. Auf dem QQ-Plot sieht man, dass die Werte von der Normalverteilungslinie abweichen. Der Shapiro-Wilk normality test hat für diesen Datensatz einen p-value = 0.04582, was die Nullhypothese (H0=Daten sind nicht normalverteilt) für das 5% Konfidenzniveau knapp verwirft. Es liegt demnach eine Normalverteilung vor mit einem Median von 53.25 als zentrales Lagemaß.

Frost: Dieser Parameter gibt die durchschnittlichen Tage unter 0 Grad Celsius zwischen 1931-1960 an. Laut Shapiro-Wilk normality test mit p-value = 0.05267, kann H0 (H0=Daten sind nicht normalverteilt) ganz knapp für das 5% Konfidenzniveau nicht verworfen werden. Auch der QQ-Plot und der Kurtosis-Wert mit -0.859056 sprechen für eine Normalverteilung. Das Histogramm lässt darauf schließen, dass es sich um eine unimodale Verteilung handelt mit einem kleinen zusätzlichen Peak im Bereich 0-25. Es empfiehlt sich der robuste Median als zentrales Lagemaß und die IQD als robustes Streumaß.

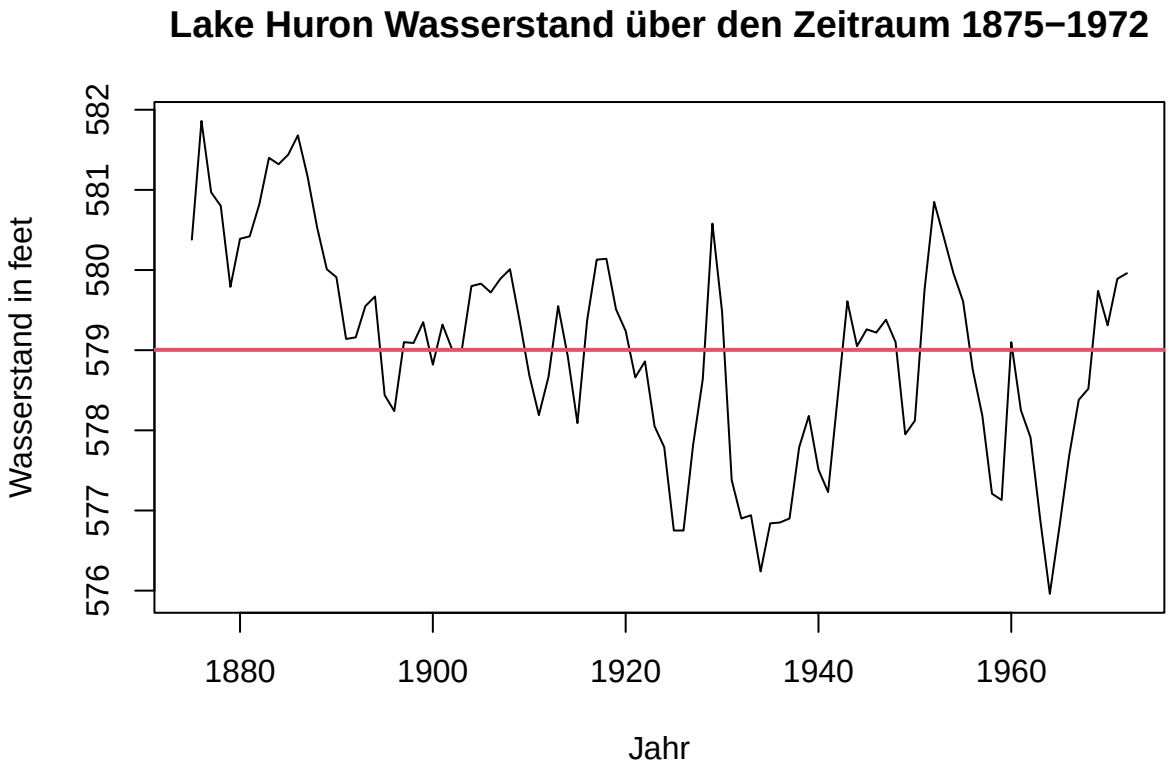
**Korrelation** In der Korrelationsmatrix sind einige der Korrelationen hochsignifikant. So korrelieren Murder und Illiteracy mit 0,70 positiv zueinander. Auch ist die Life Exp mit Murderer negativ korreliert, wobei dies vermutlich auf die generellen Lebensumstände in den jeweiligen Bundesstaaten zurückzuführen ist (eine durchschnittliche Mord-Inzidenz von 7.378/100.000 wird sich dimensionstechnisch nicht auf die Lebenserwartung auswirken). Ebenfalls korrelieren mit 0,7 Murder ~ Illiteracy positiv zueinander. Dementgegen korreliert Illiteracy mit HS Grad mit -0,66 negativ, wobei dies aufgrund des deutlichen Zusammenhangs von Schulbildung und Analphabetismus eigentlich schon als Kausalität interpretiert werden kann. Es gibt jedoch auch hochsignifikante Korrelationen mit dem Parameter Frost, die jedoch keine logischen Interpretationen zulassen. (Frost~Illiteracy = -0.67***).

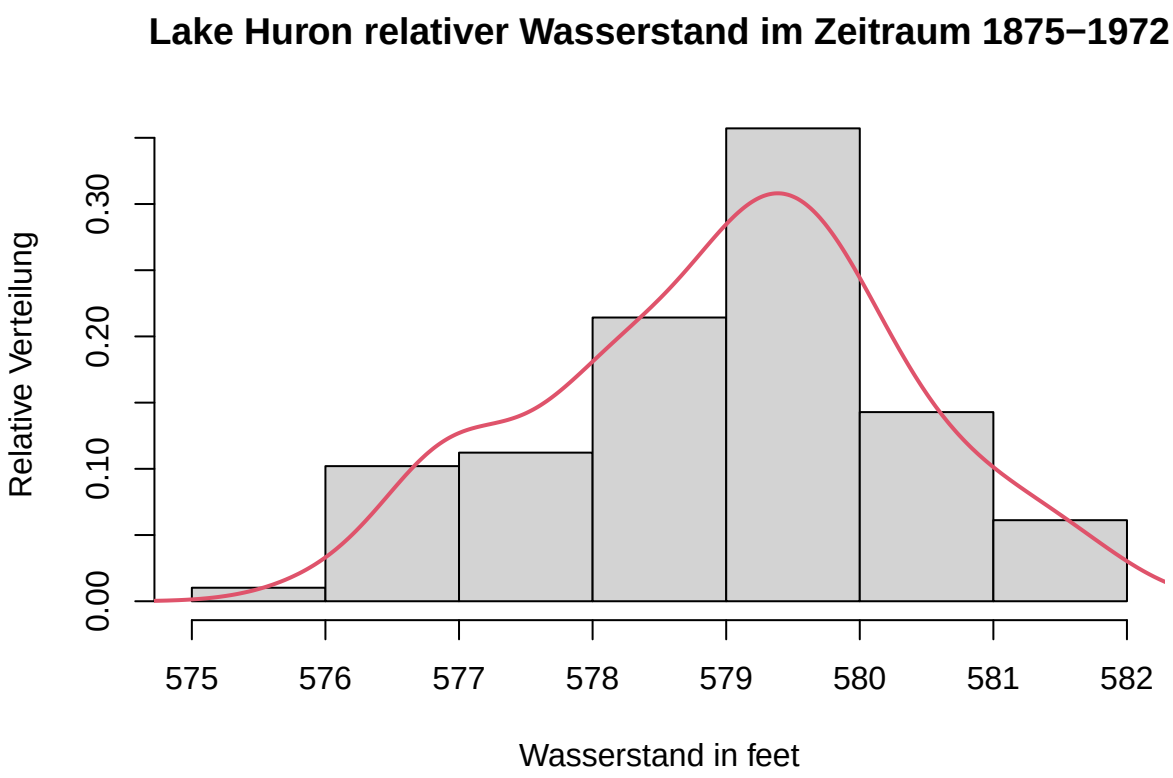
**Aufgabe 3**

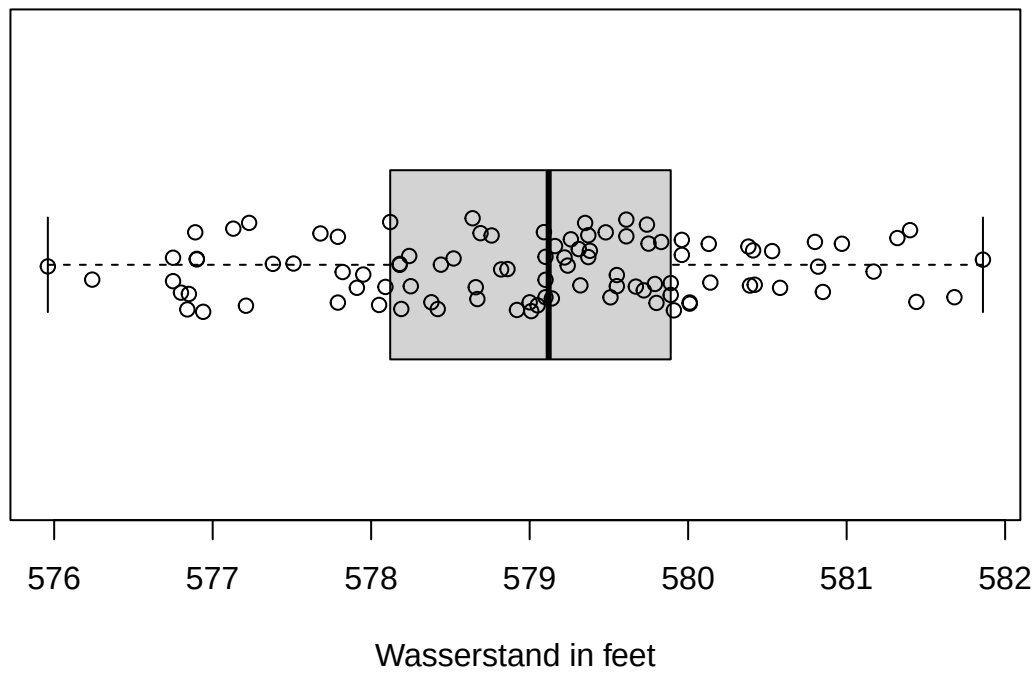
Explorieren und visualisieren Sie den Datensatz “LakeHuron” mit und ohne Berucksichtigung des Zeitreihenaspektes. • wählen Sie sinnvolle Schätzer fur Lokation, Variation, Schiefe and Gewicht in den Rändern. • Geben Sie dem Nutzer die Möglichkeit zwischen unterschiedlichen graphischen Darstellungen zu wechseln. Erklären Sie die Zusammenhänge und Eiegenschaften der Daten, die sich aus diesen Visualisierungen erkennen lassen. – Sind die Daten symmetrisch/schief? – Haben die Daten schwere Ränder? – Bieten Sie robuste und nichtrobuste Lagemaße und Skalenmaße im Vergleich oder zur Auswahl an. – Sind die Daten (approximativ) normalverteilt?

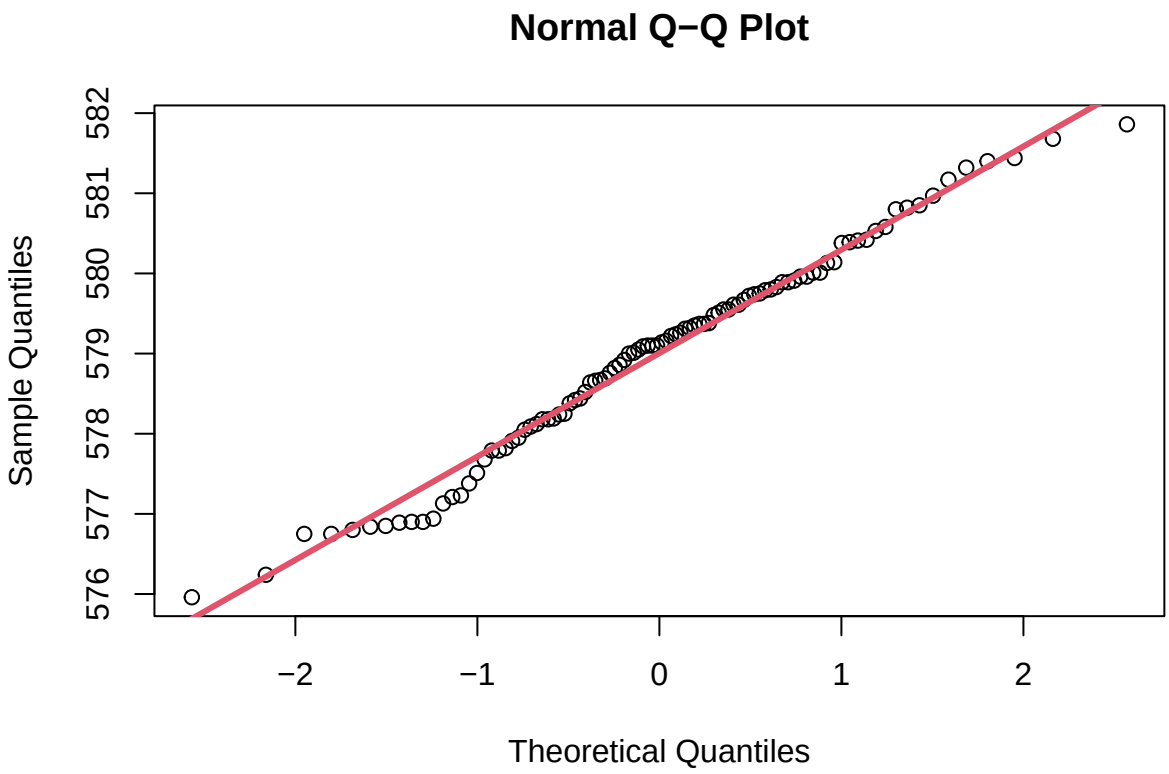
```
library(moments)
library(hrbrthemes)
```

Bei dem Datensatz “LakeHuron” handelt es sich um jährlich gemessenen Wasserstand in “feet” über den Zeitraum 1875-1972. Insgesamt enhält der Datensatz 98 Messdaten.




Lake Huron relativer Wasserstand im Zeitraum 1875−1972
Relative Verteilung
0.00
0.10
0.20
0.30
575
576
577
578
579
580
581
582
Wasserstand in feet


Lake Huron Wasserstand im Zeitraum 1875−1972
576
577
578
579
580
581
582
Wasserstand in feet




Der Wasserstand variierte im Zeitraum 1875-1972 zwischen 575.96 und 581.86, mit einem nicht-robusten Mittelwert von 579.0040816 und der Standardabweichung 1.3182985 bzw. dem robusten Median 579.12 und der IQD 1.74. Beide Lagemaße sind sehr ähnlich, deshalb konnten keine Ausreisser in diesem Datenset festgestellt werden. Auch die zwei Streumaße Standardabweichung und IQD sind von den Werten her ähnlich. Die Verteilung ist unimodal, was aufgrund eines einzelnen Datenpeak festzustellen ist.

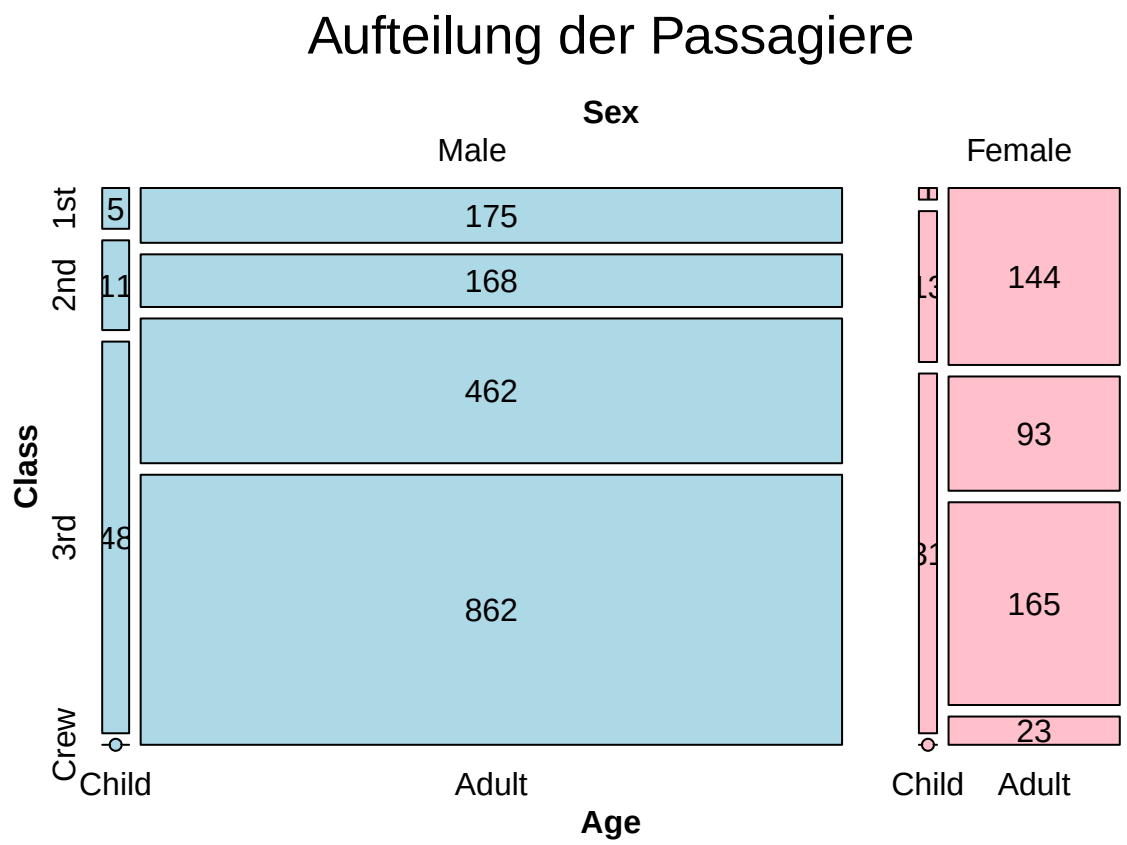
Der Skewness-Wert von -0.1397719 deutet auf eine linksschiefe Verteilung hin, was auch im Histogramm ersichtlich ist. Mit einem Kurtosis-Wert von -0.500837 kann eine flachgipflige, approximative Normalverteilung festgestellt werden, die auch im Histogramm und im QQ-Plot visuell bestätigt wird. Die Daten deuten auf einen linksseitigen, schweren Rand und einem rechtsseitigen, leichten Rand hin, was im QQ-Plot abzulesen ist.

**Aufgabe 4**

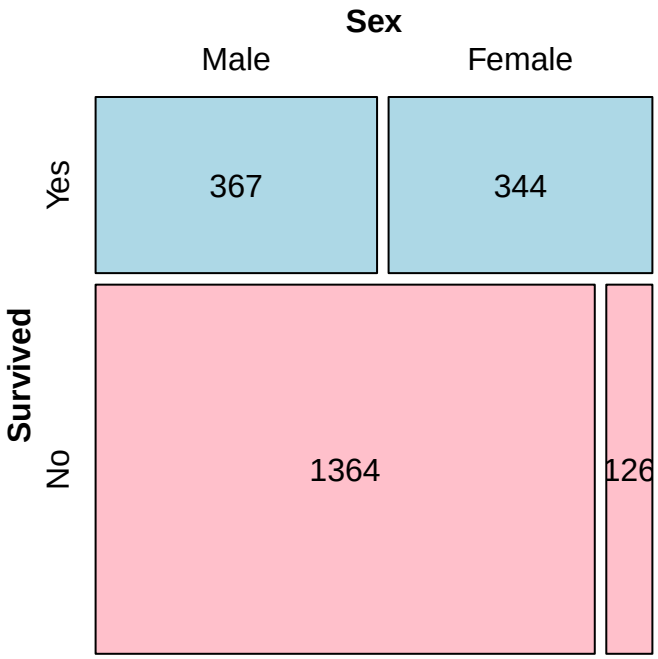
Explorieren und visualisieren Sie den Datensatz Titanic. Wie beeinflussen Alter, Geschlecht und Klasse das Überleben? Finden Sie, wo sich Simpson's Paradoxon zeigt und begründen Sie, woher dieser Effekt kommt.

```
titanic_df <- as.data.frame(Titanic)
ftable(titanic_df[1:4, 1:4])
```

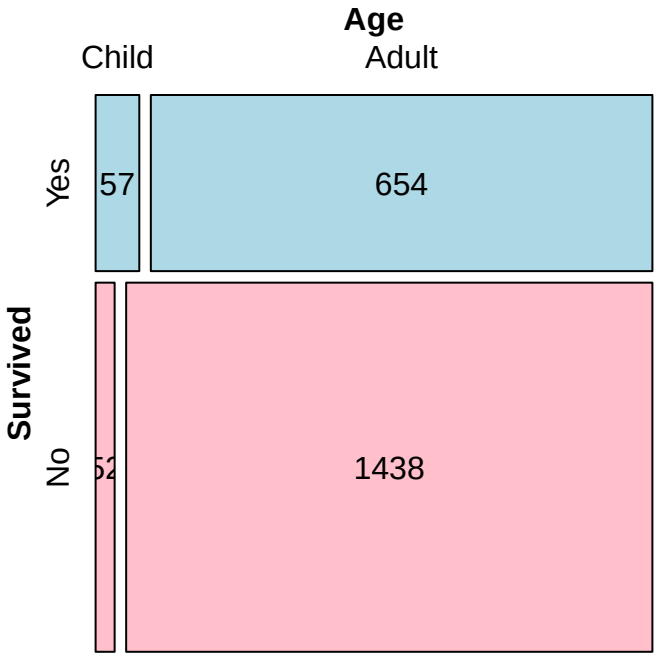
```
##                     Survived No Yes
## Class Sex    Age
## 1st   Male   Child           1   0
##              Adult           0   0
##       Female Child           0   0
##              Adult           0   0
## 2nd   Male   Child           1   0
##              Adult           0   0
##       Female Child           0   0
##              Adult           0   0
## 3rd   Male   Child           1   0
##              Adult           0   0
##       Female Child           0   0
##              Adult           0   0
## Crew  Male   Child           1   0
##              Adult           0   0
##       Female Child           0   0
##              Adult           0   0
```

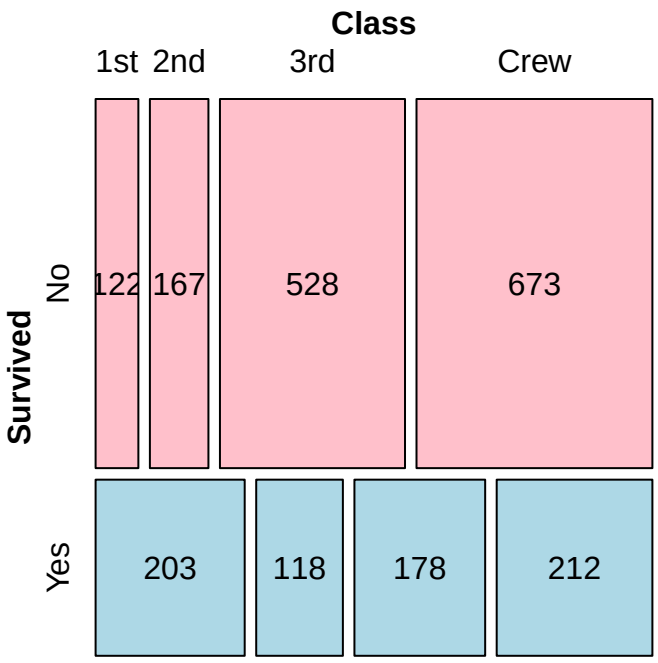


# Überleben nach geschlecht



# Überleben nach Alter






```
## [1] 0.6246154
```

```
## [1] 0.4140351
```

```
## [1] 0.2521246
```

```
## [1] 0.239548
```

Dieser Datensatz enthält Informationen über das Schicksal der Passagiere auf der tödlichen Jungfernfahrt des Ozeandampfers “Titanic”, zusammengefasst nach wirtschaftlichem Status (Klasse), Geschlecht, Alter und Überleben.

Geschlecht Man kann im Mosaicplot “Aufteilung der Passagiere” erkennen, dass die meisten Passagiere männlich sind. Von den Männern sind mit Abstand die meisten Crew-Mitglieder oder in der dritten Klasse. Sieht man sich nun das Überleben nach Geschlecht an, so überleben in absoluten Zahlen fast gleich viele Männer wie Frauen. Vom Verhältnis sind mehr als 10x soviele Männer gestorben wie überlebt haben. Bei den Frauen haben etwa doppelt soviele überlebt wie gestorben sind.

Alter Hier ist auch eine deutliche Differenz erkennbar. Etwa gleich viele Kinder haben überlebt wie gestorben sind. Bei den Erwachsenen ist das Verhältnis etwa 1:2 (Überlebt/Gestorben).

Klasse Auch die Klasse hat die Überlebenswahrscheinlichkeit deutlich beeinflusst. In der ersten Klasse haben etwa 76% der Passagiere überlebt, in der zweiten Klasse 62%, in der dritten Klasse 25% und bei der Crew mit 24% am wenigsten.

## Simpson's Paradoxon

Das Simpson Paradoxon beschreibt einen Trend in Daten, der sich jedoch auflöst oder umkehrt, wenn die einzelnen Gruppen betrachtet werden. Die Ursache ist meist eine nicht betrachtete Störvariable oder eine unterschiedliche Gewichtung der Daten. In diesem Beispiel zeigt es sich in den Überlebenswahrscheinlichkeiten der Crew und dritten Klasse. Auf den ersten Blick liegt die Überlebensrate für die dritte Klasse mit 25,21% etwas höher als jene der Crew mit 23,95%. Betrachtet man die beiden Klassen nach Geschlecht getrennt, sieht man, dass jeweils die Todesrate in der Crew höher liegt als in der Dritten Klasse:

Überlebensrate: 3rd Männer: 17,25% Crew Männer: 22,27%

3rd Frauen: 45,91% Crew Frauen: 86,95%

3rd Gesamt: 25,21% Crew Gesamt: 23,95%

Dies lässt sich dadurch erklären, dass die Crew zu 97,40% aus Männern besteht und die Frauen eine äußerst hohe Überlebenswahrscheinlichkeit haben, während die dritte Klasse zu 72,24% aus Männern besteht. Der vergleichsweise viel höhere Frauenanteil hatte jedoch mit 45,91% noch immer eine recht hohe Überlebensrate, was den Gesamtschnitt nach oben ziehen konnte, während die 86,95% Überlebensrate der Crew-Frauen praktisch keine Auswirkung auf die Gesamtüberlebensrate der Crew hatte.

```
ftable(Titanic)
```

```
##                     Survived  No Yes
## Class Sex    Age
## 1st   Male   Child            0   5
##              Adult          118  57
##       Female Child            0   1
##              Adult            4 140
## 2nd   Male   Child            0  11
##              Adult          154  14
##       Female Child            0  13
##              Adult           13  80
## 3rd   Male   Child           35  13
##              Adult          387  75
##       Female Child           17  14
##              Adult           89  76
## Crew  Male   Child            0   0
##              Adult          670 192
##       Female Child            0   0
##              Adult            3  20
```

```
crewMAlive <- 192
crewMDead <- 670
crewMT <- crewMAlive + crewMDead

crewFAlive <- 20
crewFDead <- 3
crewFT <- crewFAlive + crewFDead

crewDead <- crewMDead + crewFDead
crewAlive <- crewMAlive + crewFAlive

thirdMAlive <- 13+75
thirdMDead <- 35+387
```

```r
thirdMT <- thirdMAlive + thirdMDead

thirdFAlive <- 14+76
thirdFDead <- 17+89
thirdFT <- thirdFAlive + thirdFDead

thirdAlive <- thirdMAlive + thirdFAlive
thirdDead <- thirdMDead + thirdFDead

crewMT; crewFT;  crewMT + crewFT
```

```
## [1] 862
```

```
## [1] 23
```

```
## [1] 885
```

```r
crewDead; crewAlive
```

```
## [1] 673
```

```
## [1] 212
```

```r
thirdMT; thirdFT; thirdFT+ thirdMT
```

```
## [1] 510
```

```
## [1] 196
```

```
## [1] 706
```

```r
thirdDead; thirdAlive
```

```
## [1] 528
```

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

```r
#Survival Rates

#total
ThirdSurvRate <- thirdAlive/(thirdAlive+thirdDead); ThirdSurvRate*100
```

```
## [1] 25.21246
```

```r
CrewSurvRate <- crewAlive/(crewAlive+crewDead); CrewSurvRate*100
```

```
## [1] 23.9548
```

```r
#third class sex
ThirdMSurvRate <- thirdMAlive/(thirdMAlive+thirdMDead); ThirdMSurvRate*100
```

```
## [1] 17.2549
```

```r
ThirdFSurvRate <- thirdFAlive/(thirdFAlive+thirdFDead); ThirdFSurvRate*100
```

```
## [1] 45.91837
```

```r
#crew sex
CrewMSurvRate <- crewMAlive/(crewMAlive+crewMDead); CrewMSurvRate*100
```

```
## [1] 22.27378
```

```r
CrewFSurvRate <- crewFAlive/(crewFAlive+crewFDead); CrewFSurvRate*100
```

```
## [1] 86.95652
```

```r
#Geschlechterverhältnis
CrewVerhältnis <- crewMT/(crewMT+crewFT); CrewVerhältnis*100
```

```
## [1] 97.40113
```

```r
thirdVerhältnis <- thirdMT/(thirdMT+thirdFT); thirdVerhältnis*100
```

```
## [1] 72.23796
```