HAUSÜBUNG 2

Explorative Datenanalyse, Erkennen von Eigenschaften und Visualisierung von unterschiedlichen Variablentypen

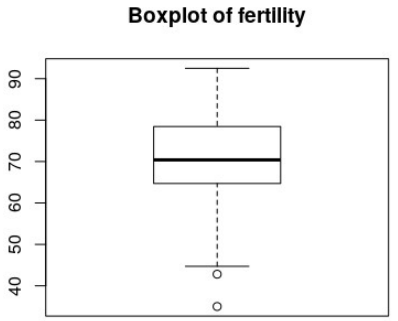
Aufgabe 1

Bei den Daten handelt es sich um eine standardisierte Fruchtbarkeitsmessung mit gemeinwirtschaftlichen Faktoren aller französischsprachigen Schweizer Provinzen (47) aus 1888.

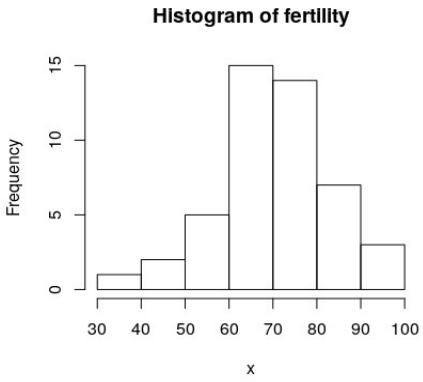
Die Variablen *Fertility* (in %), *Agriculture* (% der in Landwirtschaft tätigen Männer), *Education* (% der Probanden die Ausbildung weiter als normale Schulreife haben), *Catholic* (% der Probanden die katholisch sind) und *Infant Mortality* (Lebendgeburten die weniger als ein Jahr überleben).

Fertility:

Lokation: Der arithmetische Mittelwert der Daten liegt bei 70,14%, der Median bei 70,40% und die Daten haben eine Interquantilsdistanz von 13,75%. Es wurden im Boxplot zwei Ausreißer gefunden.

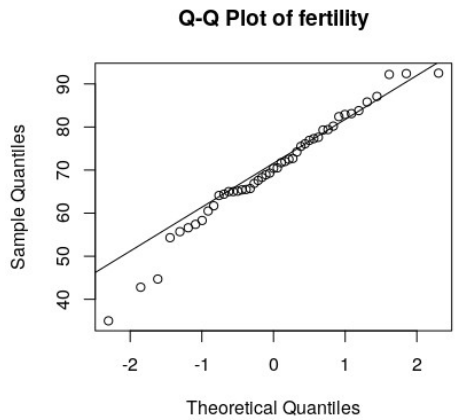


Die



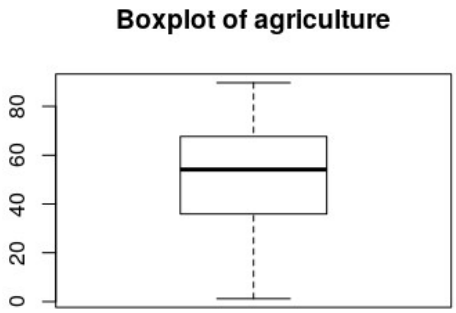
Standardabweichung beträgt 12,49% bei einer mittleren Abweichung vom Median (MAD) von 10,23%.

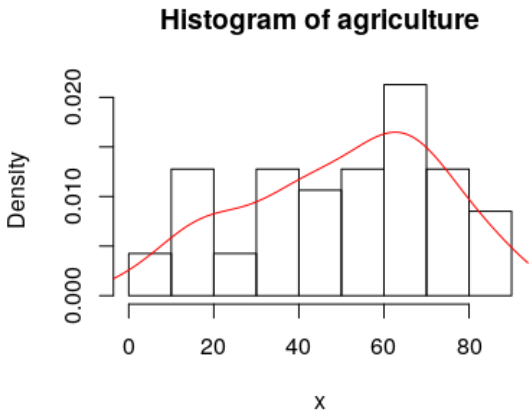
Schiefe und Gewicht in den Rändern: Die Daten sind symmetrisch, unimodal und haben eine berechnete Schiefe von -0,46, sie sind also leicht linksschief. Sie weisen wie im Q-Q Plot ersichtlich links einen schweren Rand auf (die Daten streuen sich weiter vom Median).



Agriculture:

Lokation: Die Agriculture Daten weisen einen arith. Mittelwert von 50,65% und einen Median von 54,1% auf. Ausreißer gibt es keine.





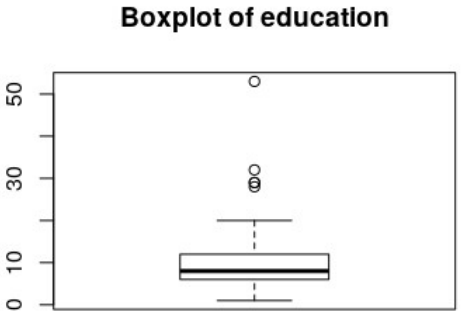
Die Standardabweichung beträgt 22,71% und ist damit relativ hoch. Sie ist bei einem Variationskoeffizienten von 0,45 knapp die Hälfte des Mittelwerts und die MAD ist 23,87%.

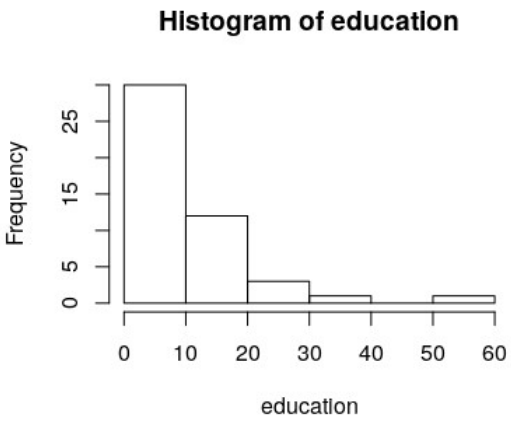
Schiefe und Gewicht in den Rändern: Die Daten sind nicht normalverteilt und zeigen im Histogramm Anzeichen von Multimodalität. Allerdings sind die “Peaks” bei Betrachtung der Dichtefunktion nicht mehr eindeutig und auch von einer Gleichverteilung kann durch den ungleichen Verlauf kaum gesprochen werden. Im Q-Q Plot erkennt man rechts einen leichten Rand.

Bei einer berechneten Schiefe von -0,32 wären die Daten insgesamt leicht linksschief, allerdings ist es schwer bei diesem Datensatz konkretere Aussagen zu treffen.

Education:

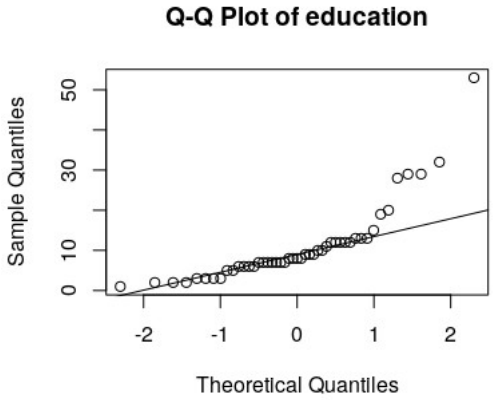
Lokation: Die Daten weisen einen Mittelwert von 10,98% und einen Median von 8% auf. Im Boxplot werden insgesamt 5 Werte als Ausreißer erkannt, wobei 4 davon noch unter 35% liegen und einer über 50%.



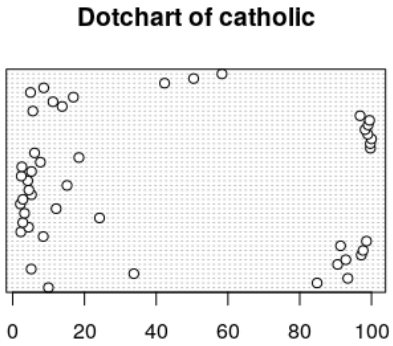


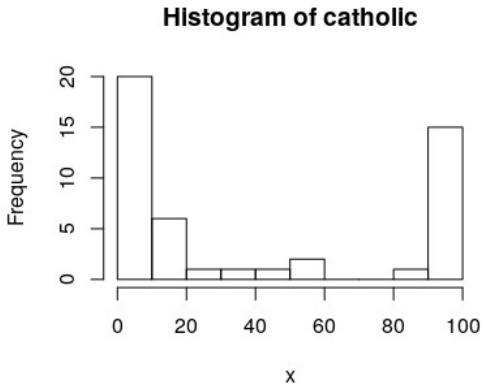
Die Standardbweichung beträgt 9,62% und die MAD 5,93%.

Schiefe und Gewicht in den Rändern: Die Daten sind stark rechtsschief mit einer berechneten Schiefe von 2,27. Sie sind unimodal und weisen einen schweren Rand an der rechten Seite auf.



Catholic:

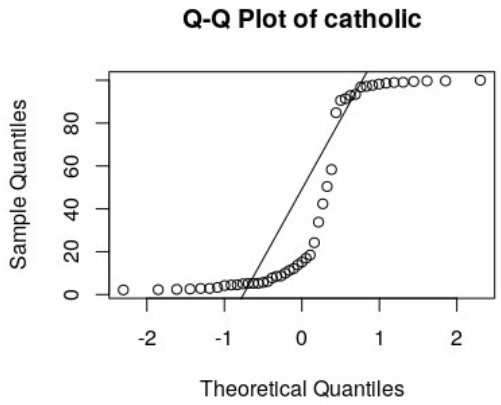




Lokation: Durch die starke Bimodalität dieser Daten ist der arithmetische MW kein geeigneter Schätzer und auch der Boxplot wird hier nicht herangezogen. Der Median beträgt 15,14%, das 1. Quantil 5,19% und das 3. Quantil 93,13%, also eine IQD von fast 88%. Wie aus dem Dotchart und dem Histogramm zu erkennen ist, sammeln sich die meisten Daten entweder an einem Extrem (niedriger Anteil an katholisch orientierten) oder dem anderen (sehr viele Katholiken).

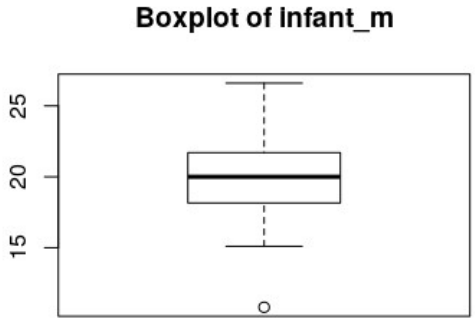
Durch die Bimodalität lassen sich keine sinnvollen Aussagen über Varianz und Schiefe treffen.

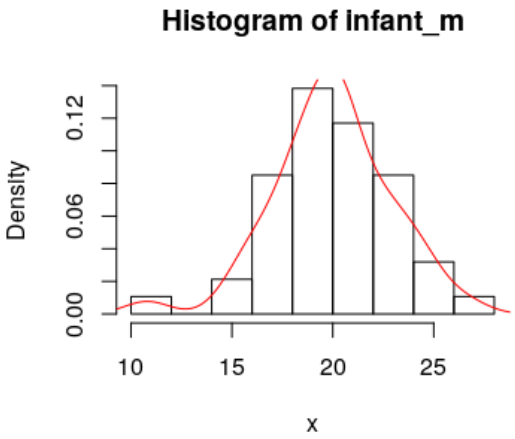
Auch wenn es in diesem Datensatz klar ist, dass es sich um keine Normalverteilung handelt sieht man anhand des Q-Q Plots gut die typische S-Formung von bimodalen Daten.



Infant Mortality:

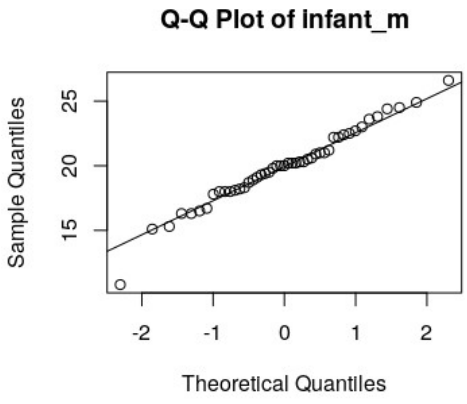
Lokation: Dieser Datensatz weist einen arithmetischen Mittelwert von 19,94% und einen Median von 20% aus, es wurde ein Ausreißer gefunden.





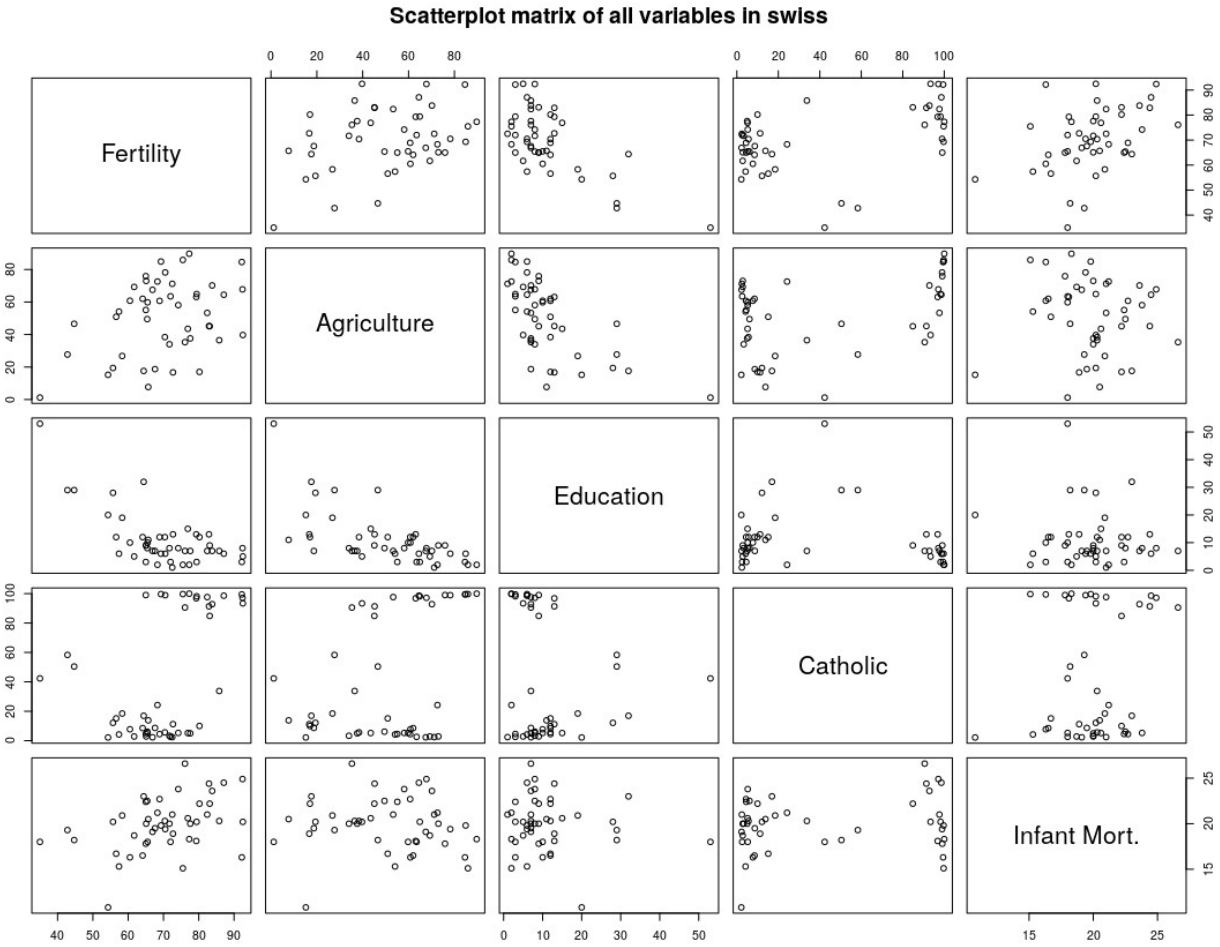
Die Standardabweichung ist mit 2,91% relativ gering.

Schiefe und Gewicht in den Rändern: Die Daten Infant Mortality sind annähernd normalverteilt. Sie sind mit einer berechneten Schiefe von -0,33 leicht linksschief (dies lässt sich vermutlich auf den Ausreißer zurückführen) und weisen keine schweren oder leichten Ränder auf (lediglich ganz links im QQ Plot geht der Ausreißer vom Zentrum weg).



Zusammenhänge zwischen den Variablen:

Zur übersichtlichen Darstellung möglicher Zusammenhänge sei eine Scatterplot Matrix angeführt:



Auf ersten Blick ist es sehr schwierig bei diesen Daten einen Zusammenhang zu erkennen. Durch Berechnen der Korrelationsmatrix (nach Spearman) ergibt sich eine übersichtlichere Tabelle:

	Fertility	Agriculture	Education	Catholic	Infant.M
Fertility	1.0000000	0.2426643	-0.44325769	0.41364556	0.43713670
Agriculture	0.2426643	1.0000000	-0.65046381	0.28868781	-0.15212866
Education	-0.4432577	-0.6504638	1.00000000	-0.14441631	-0.01898137
Catholic	0.4136456	0.2886878	-0.14441631	1.00000000	0.06611714
Infant.M	0.4371367	-0.1521287	-0.01898137	0.06611714	1.00000000

Aus dieser Tabelle ist besser ersichtlich, dass etwa die Daten Fertility und Agriculture mit Education negativ korrelieren (-0,44 bzw. -0,65) oder etwa Fertility und Catholic bzw. Infant Mortality positiv (0,41 bzw. 0,44).

Diese Korrelationskoeffizienten entsprechen alle etwa einer “medium correlation”, wobei Fertility und Education als einzige einen Wert über 0,5 aufweisen und man am ehesten bei diesen Daten von einer Korrelation ausgehen kann. Man kann also sagen, dass bei steigendem Ausbildungsniveau die Fruchtbarkeitsraten sinken und zwischen diesen beiden Variablen der stärkste Zusammenhang besteht.

Aufgabe 2

Der Datensatz zeigt die 50 Staaen von Amerika in der ersten Spalte; die weiteren Spalten beinhalten Daten zu Population, Income, Illiteracy, Life expectation, Murder, HS Grad, Frost, Area. Insgesamt gibt es 8 Spalten.

→ Datensatz in R als x speichern: x <-state.x77

Population (pop = x[,1])

```
> summary(pop)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    365    1080    2838    4246    4968   21198
```

Da sich die Werte des Mittelwerts und Medians unterscheiden und beide Werte nicht ziemlich vorne in der range der Daten liegen ist eine weitere Exploration nötig. Man könnte derzeit eventuell davon ausgehen, dass die Daten in eine Richtung mehr verzerrt sind.

1) plot(pop)

→ der Plot zeigt, dass die meisten Daten im Bereich 0-5000 liegen

2) hist(pop, freq=F, breaks=5, main=paste("Population histogram of all 50 states in America"))

lines(density(pop), col=2)

→ das Histogramm zeigt deutlich, dass die Daten im Bereich 0-5000 liegen, anhand der density line kann man erkennen, dass es noch einen zweiten sehr leichten peak bei ca. 12000 gibt, es handelt sich aber eher um eine unimodale Verteilung (d.h. man kann einen Lageschätzer verwenden), die in diesem Fall rechts schief ist, somit ist das erste Kriterium der Normalverteilung gebrochen (Daten sind nicht symmetrisch)

3) boxplot(pop, horizontal=TRUE)

→ man sieht auch beim boxplot, dass 75% der daten im Bereich 0-5000 liegen, die restlichen 25% gehen bis auf10000. Weiters gibt es 5 Ausreißer Datenpunkte. Wenn 5 von 100 Datenpunkten deutlich Ausreißer sind muss man an Normalverteilung zweifeln. Es ist

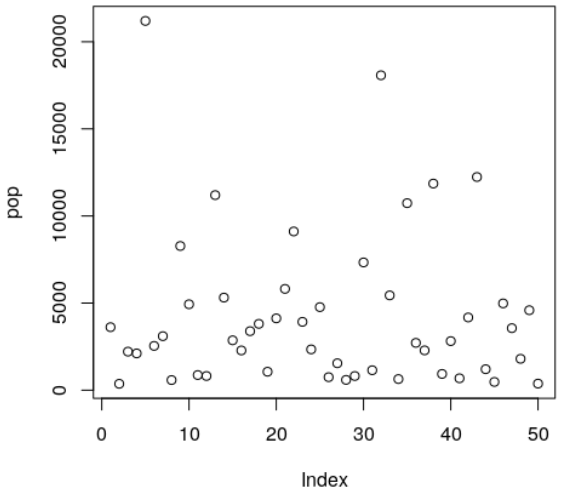
deutlich sichtbar, dass die Daten vermehrt zwischen 0-500 liegen, deshalb kann man eine Normalverteilung ausschließen.

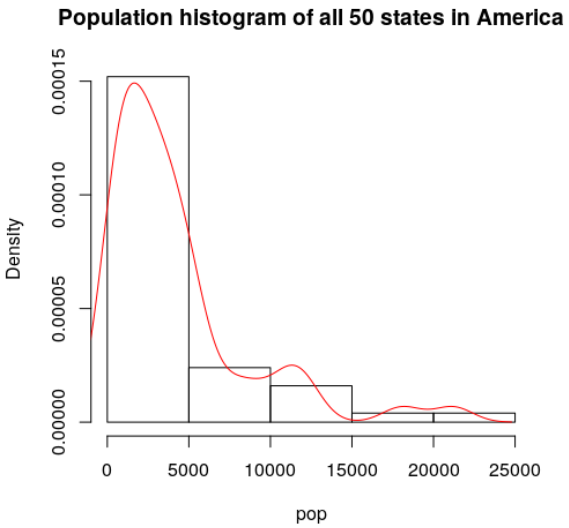
4) qqnorm(pop); qqline(pop)

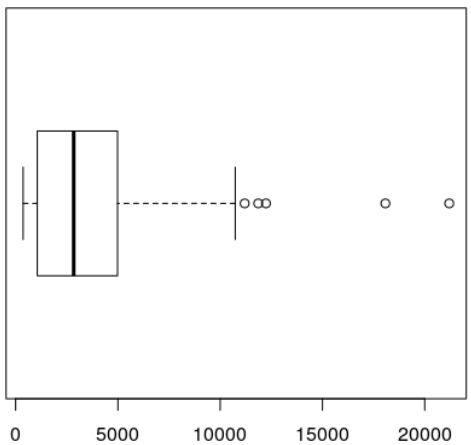
→ hier sieht man, dass es einmal einen schweren Rand (rechts vom median) und einmal einen

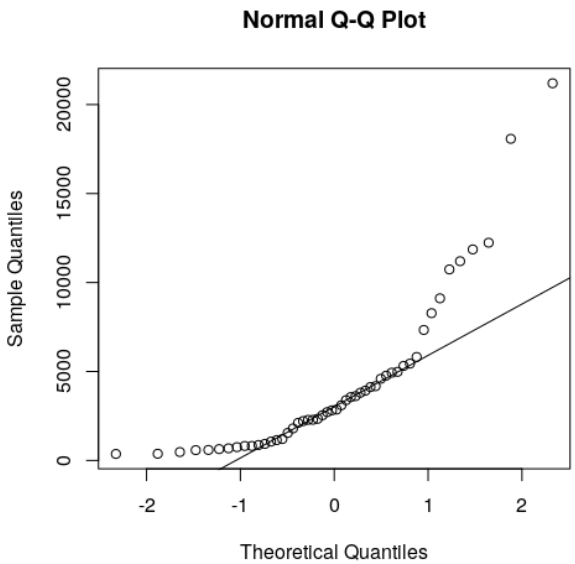
leichten Rand (links vom median) gibt. Deshalb handelt es sich eindeutig um eine rechts schiefe Verteilung und die Normalverteilung kann ausgeschlossen werden.

Der Median wird in diesem Fall als Lageschätzer verwendet werden, da er der beste Lageschätzer ist, wenn die Daten nicht normalverteilt sind und er ist robust gegenüber Ausreißern. Als Streuungsmaß kann man den Interquartilsabstand (IQR = 3889) und als Lagemaß den median absolute deviation (MAD) (mad = 2890,329) verweden.









Income (inc = x[,2])

```
> summary(inc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3098    3993    4519    4436    4814    6315
```

Der Mean und der Median sind ziemlich gleich und liegen auch im mittleren Bereich zwischen

maximalem und minimalem Wert.

1) plot(inc)

→ zeigt, dass Werte verteilt vermehrt im Bereich zwischen 3500 und 5500 liegen

2) hist(inc, freq=F, breaks=5, main=paste("Income histogram of all 50 states in America"))

lines(density(inc), col=2)

→ das Histogramm zeigt, dass es sich um eine unimodale Verteilung handelt, somit kann man den Boxplot anwenden

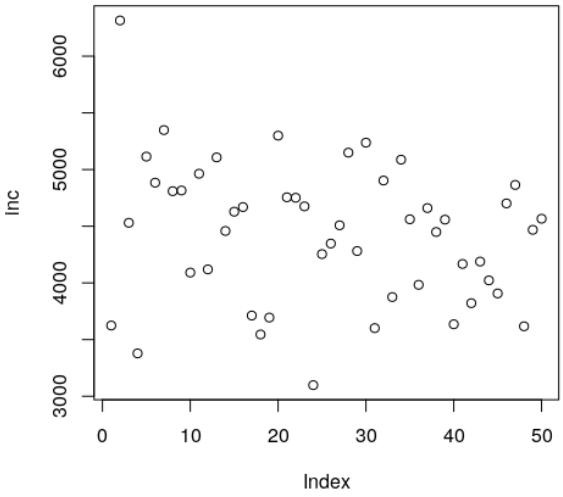
3) boxplot(inc, horizontal=TRUE)

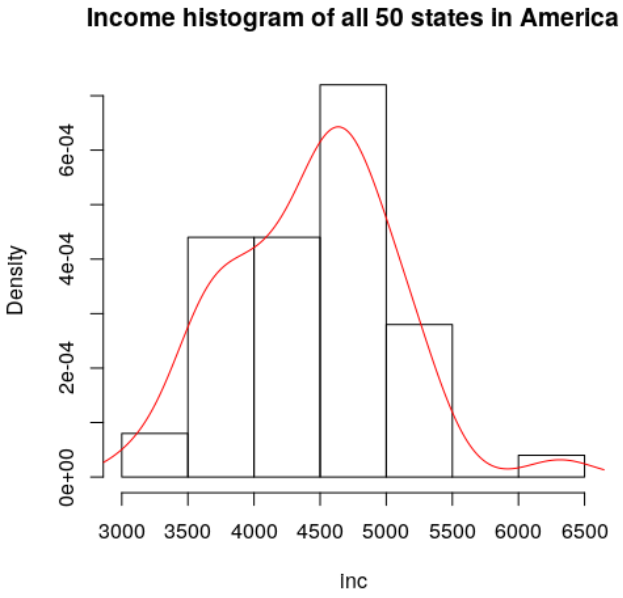
→ Boxplot zeigt, dass es einen Ausreißer gibt, und dass die mittleren 50% der Daten eher rechts liegen

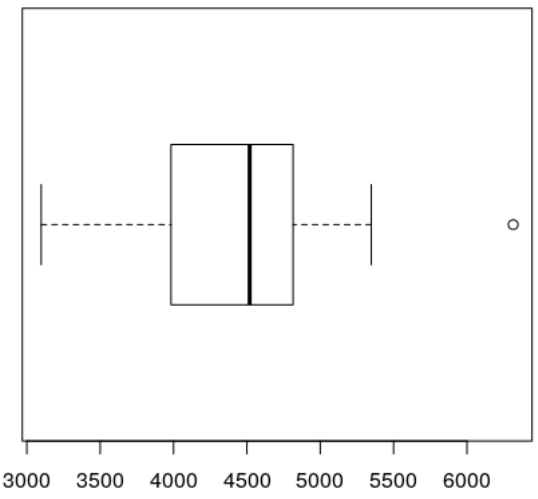
4) qqnorm(inc); qqline(inc)

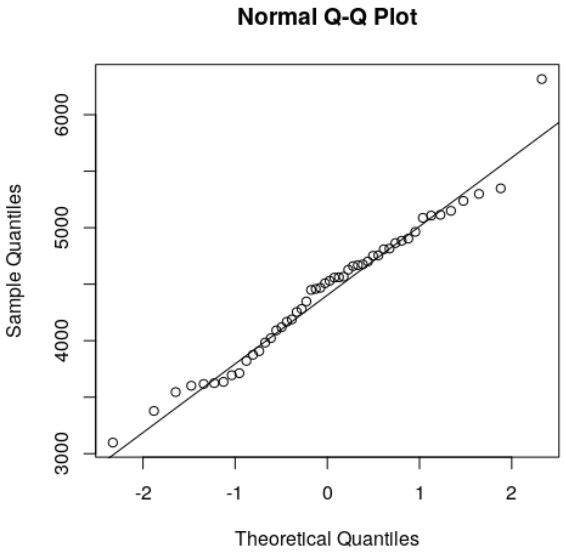
→ der qqplot zeigt, dass die Daten deutlich auf der Geraden liegen und nur ganz leicht abweichen, man sieht leichte Ränder auf beiden Seiten

Die Daten sind annäherend normalverteilt, da die Daten symmetrisch sind und es sich nur um leichte Ränder und keine schweren Ränder handelt. Außerdem sind die Daten unimodal. In diesem fall wendet man den Mittelwert und die Standardabweichung an. (SD = +/- 614,47)


Inc
6000
5000
4000
3000
0
10
20
30
40
50
Index


Income histogram of all 50 states in America
Density
6e-04
4e-04
2e-04
0e+00
3000
3500
4000
4500
5000
5500
6000
6500
Inc


3000
3500
4000
4500
5000
5500
6000


Normal Q-Q Plot
Sample Quantiles
6000
5000
4000
3000
-2
-1
0
1
2
Theoretical Quantiles

Illiteracy (ill = x[,3])

```
> summary(ill)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.500   0.625   0.950   1.170   1.575   2.800
```

Der Mean und der Median sind halbwegs ähnlich jedoch liegen sie nicht relativ in der Mitte zwischen min und max Wert. Max Wert deutlich höher.

1) plot(ill)

→ der plot zeigt, dass die Daten wild verstreut sind weiter Analyse notwenig

2) hist(ill, freq=F, breaks=5, main=paste("Illiteracy histogram of all 50 states in America"))

lines(density(ill), col=2)

→ die Daten sind unimodal, und rechts schief, die Daten liegen vermehrt im 0,5-1,5 Bereich, damit ist die Symmetrie gebrochen und man kann nicht von einer Normalverteilung ausgehen. Außerdem ist eine leichte zweiter Erhöhung in der densitiy line sichtbar.

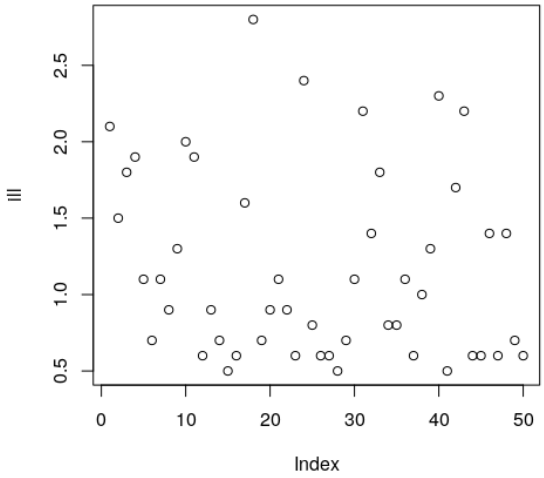
3)boxplot(ill, horizontal=TRUE)

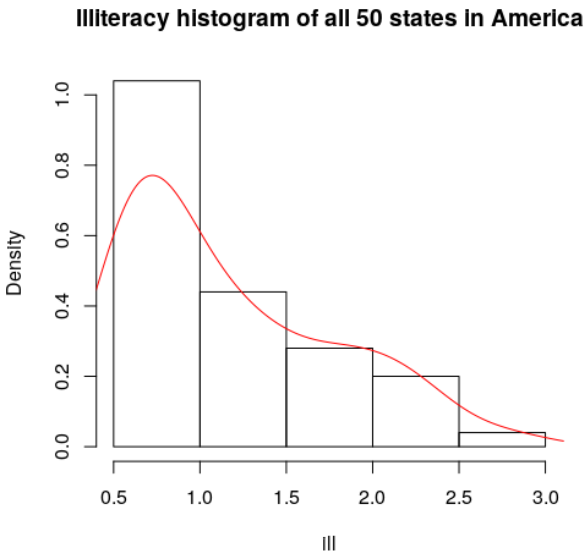
→ auch hier sieht man, dass die äußeren linken 25% der Daten sehr breit verstreut sind was auf eine unsymmetrische Verteilung hindeutet

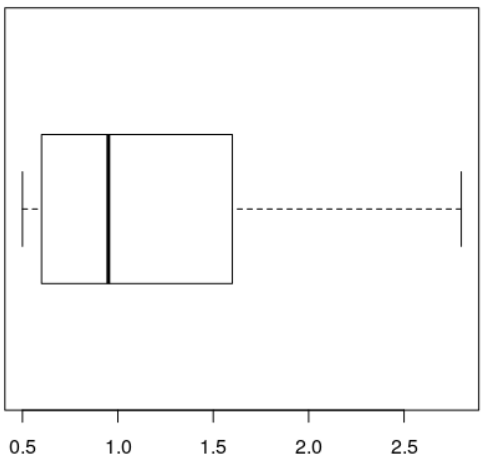
4) qqnorm(ill); qqline(ill)

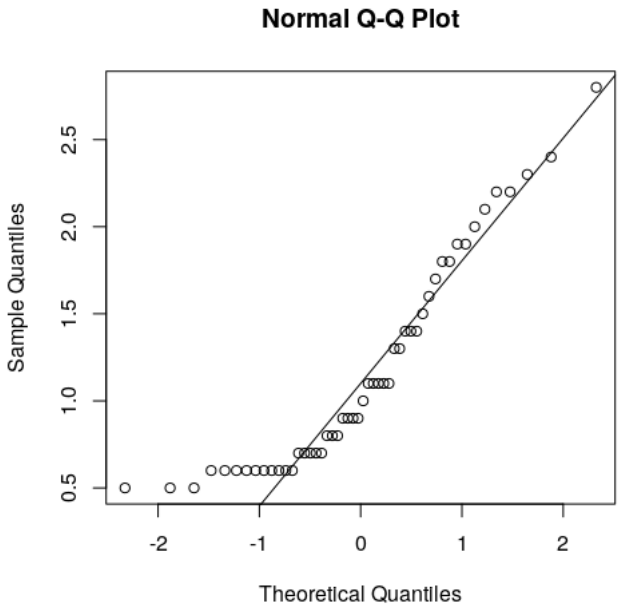
→ Daten liegen nicht auf der Geraden, und bilden eine leichte s-form was auf Multimadalität hindeutet. Zudem sieht man einen leichten Rand links vom median.

Es handelt sich um keine Normalverteilung, da hier keine symmetrie der Daten vorliegt. Die Daten sind rechts schief und der qqplot weist eine leichte s-form auf. Der Median wird in diesem Fall als Lageschätzer verwendet werden, da er der beste Lageschätzer ist, wenn die Daten nicht normalverteilt sind und er ist robust gegenüber Ausreißern. Als Streuungsmaß kann man den Interquartilsabstand (IQR = 0,95) und als Lagemaß den median absolute deviation (MAD) (mad = 0,51891) verweden.


ill
Index


Illiteracy histogram of all 50 states in America
Density
ill




Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

Life expectation (liv = x[,4])

```
> summary(liv)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  67.96   70.12   70.67   70.88   71.89   73.60
```

Mean und Median sind im gleichen Bereich und liegen ziemlich in der Mitte zwischen minimalem und maximalem Wert.

1) plot(liv)

→ man sieht, dass die Datenpunkte ziemlich verstreut sind

2) hist(liv, freq=F, breaks=5, main=paste("Life expectation histogram of all 50 states in America"))

lines(density(liv), col=2)

→ das Histogramm zeigt mit der density line eine leichten zweiten Peak und eine breitere Verteilung auf der rechten Seite

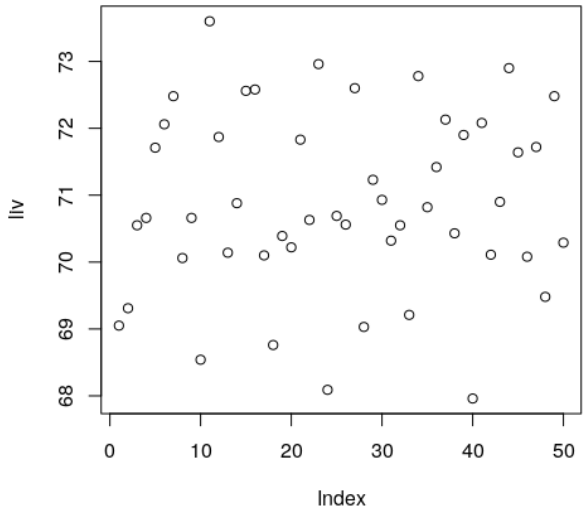
3) boxplot(liv, horizontal=TRUE)

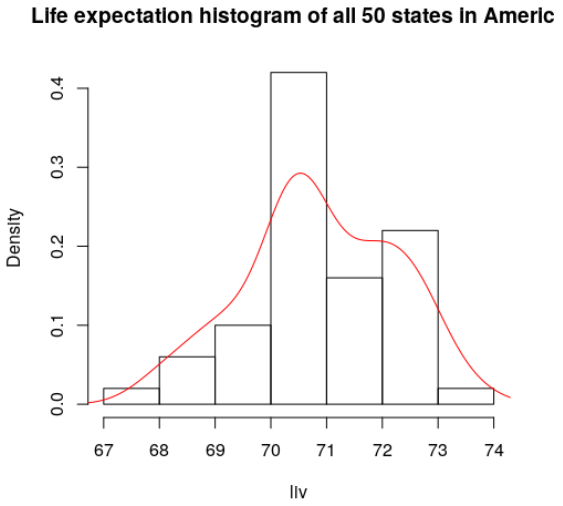
→ zeigt auch, dass die mittleren 50 prozent eher nach rechts verschoben sind

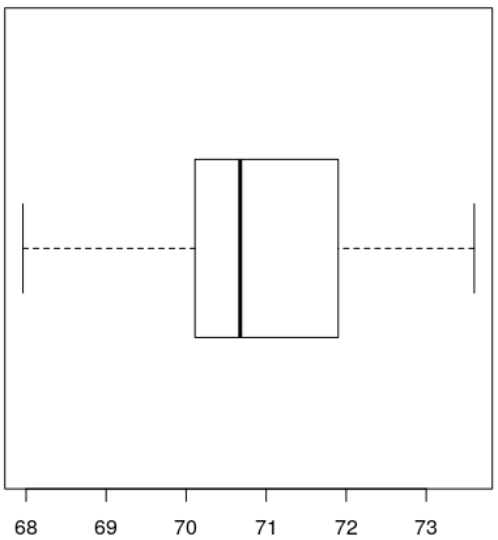
4) qqnorm(liv); qqline(liv)

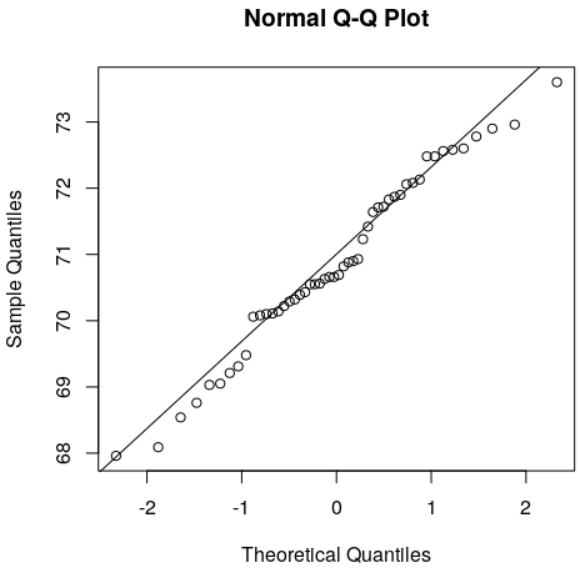
→ Daten liegen nicht unbedingt direkt auf der Geraden es bildet sich eine leichte S-form, was ein wenig auf Bimodalität hinweißt, in diesem Fall ist der boxplot keine gute analyse, außerdem sieht man einen schweren (links vom median) und einen leichten rand (rechts vom median) und was in diesem Fall auf echt schief bzw. links schief hinweist. Da hier ein schwerer Rand vorliegt muss man eine Normalverteilung ausschließen.

Der Median wird in diesem Fall als Lageschätzer verwendet werden, da er der beste Lageschätzer ist, wenn die Daten nicht normalverteilt sind und er ist robust gegenüber Ausreißern. Als Streuungsmaß kann man den Interquartilsabstand (IQR = 1.775) und als Lagemaß den median absolute deviation (MAD) (mad = 1.5419) verweden.


liv
Index
68
69
70
71
72
73
0
10
20
30
40
50


Life expectation histogram of all 50 states in Americ
Density
liv
0.0
0.1
0.2
0.3
0.4
67
68
69
70
71
72
73
74


68
69
70
71
72
73


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
68
69
70
71
72
73
-2
-1
0
1
2

Murder (mur = x[,5])

```
> summary(mur)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.400   4.350   6.850   7.378  10.675  15.100
```

Der Mean und der Median sind halbwegs ähnlich jedoch liegen sie nicht relativ in der Mitte zwischen min und max Wert. Max Wert deutlich höher. Was auf keine symmetrische verteilung hindeutet.

1) hist(mur, freq=F, breaks=5, main=paste("Murder histogram of all 50 states in America"))

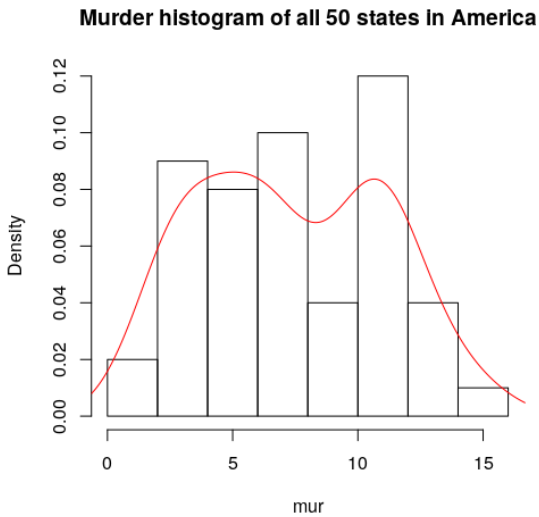
lines(density(mur), col=2)

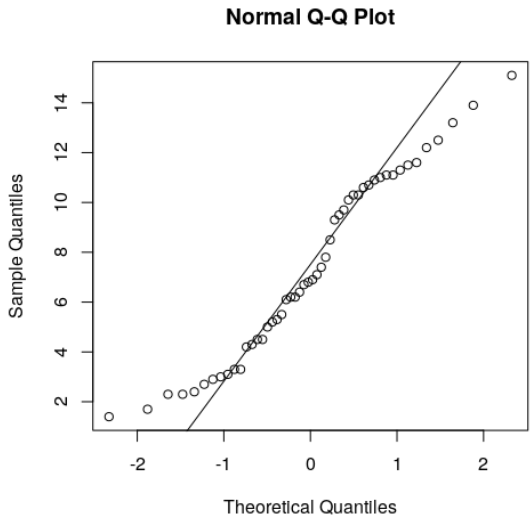
→ Daten sind deutlich bimodal, es handelt sich somit umd keine Normalverteilung, Außerdem sind die die Verteilung auch assymetrisch, bei Bimodalität kann der Boxplot nicht angewendet werden

2) qqnorm(mur); qqline(mur)

→ man sieht deutlich eine s form die auf bimodalität deutet, sowie leichte Ränder an beiden Seiten.

Keine Normalverteilung. Bei einer bimodalen Verteilung sind Lagemaße problematisch. Es könnte der Modus angewendet werden aber dieser ist nicht eindeutig.





HS Grad (grad = x[,6])

```
> summary(grad)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  37.80   48.05   53.25   53.11   59.15   67.30
```

Mean und Median sind im gleichen Bereich und liegen ziemlich zentral zwischen minimalem und maximalem Wert.

1) hist(grad, freq=F, breaks=5, main=paste("HS Grad histogram of all 50 states in America"))

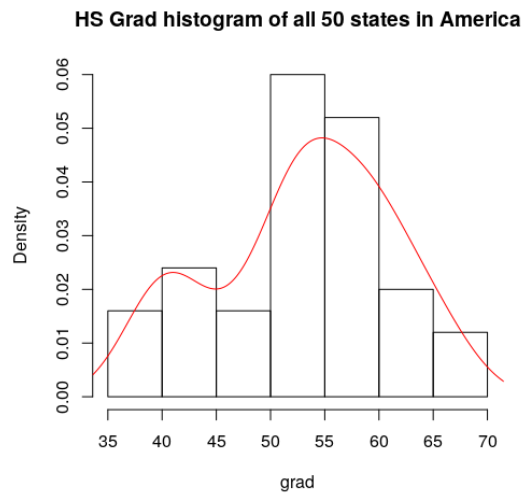
lines(density(grad), col=2)

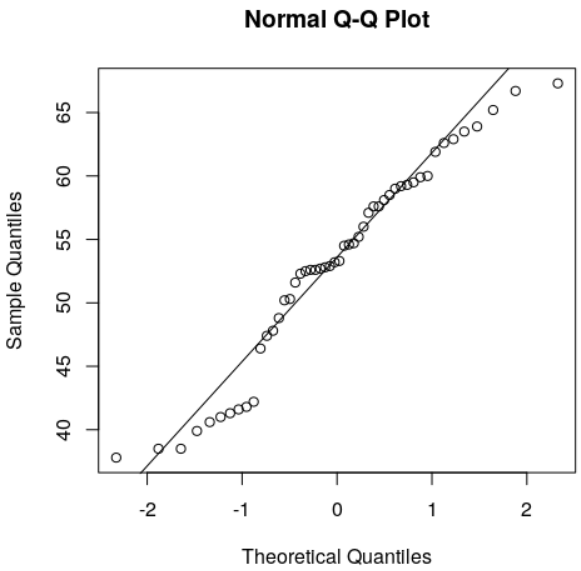
→ man sieht anhand der density line eine leichte zweite Erhöhung auf der linken Seite, außerdem sieht man eine leichte assymtrie die in den rechten Bereich der Verteilung geht

2) qqnorm(grad); qqline(grad)

→ Datenpunkte nicht auf der Geraden, außerdem sieht man einen schweren Rand (links vom median) und einen leichten Rand (rechts vom median)

Es handelt sich um keine Normalverteilung da eindeutig ein schwerer Rand im qqplot erkennbar ist. Der Median wird in diesem Fall als Lageschätzer verwendet werden, da er der beste Lageschätzer ist, wenn die Daten nicht normalverteilt sind. Als Streuungsmaß kann man den Interquartilsabstand (IQR = 11.1) und als Lagemaß den median absolute deviation (MAD) (mad = 8.6) verweden.





Frost (fro = x[,7])

```
> summary(fro)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   66.25  114.50  104.46  139.75  188.00
```

Der Mean und der Median sind halbwegs ähnlich jedoch liegen sie nicht in der Mitte zwischen min und max Wert.

1) hist(fro, freq=F, breaks=5, main=paste("Frost histogram of all 50 states in America"))

lines(density(fro), col=2)

→ Ränder sind ziemlich breit

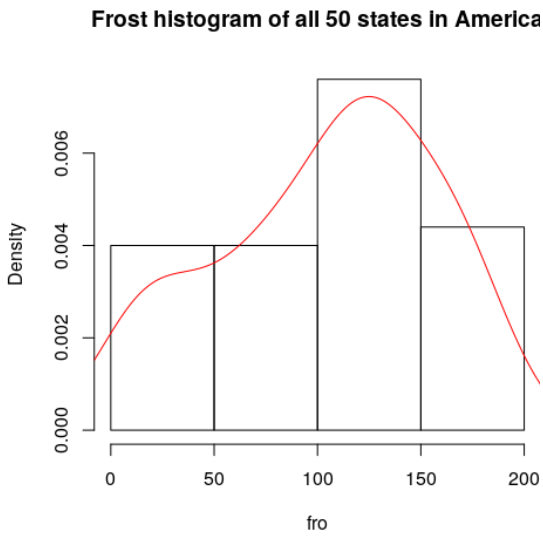
2) boxplot(fro, horizontal=TRUE)

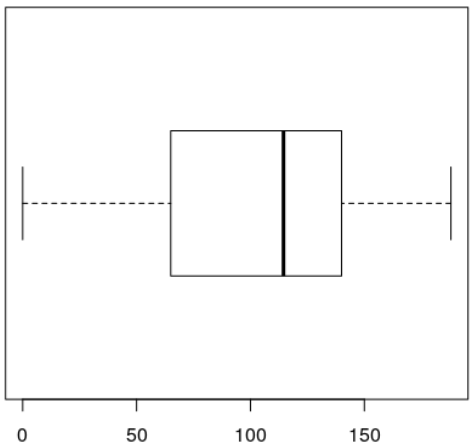
→ der Boxplot zeigt, dass die mittleren 50% mehr auf der linken Seite liegen

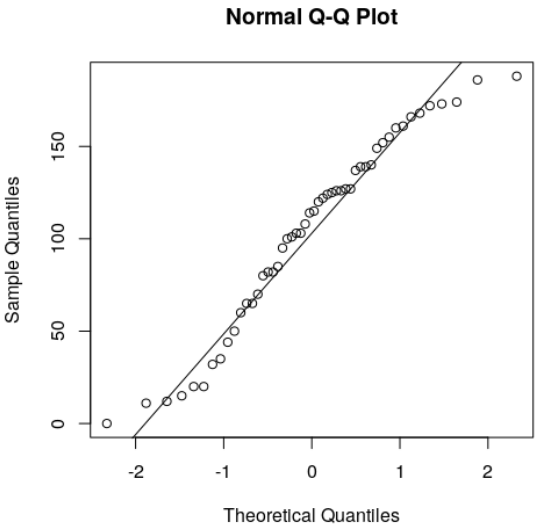
3) qqnorm(fro); qqline(fro)

→ der qqplot weist leichte Ränder an beiden Seiten auf.

Da es sich um keine schweren Ränder handelt und die Daten halbwegs symmetrisch und unimodal sind kann man sagen, dass die Daten approximativ normalverteilt sind. In diesem Fall kann man den Mittelwert und die Standardabweichung anwenden. (SD = +/- 51.98)


Frost histogram of all 50 states in America
Density
0.000
0.002
0.004
0.006
0
50
100
150
200
fro


0
50
100
150


Normal Q-Q Plot
Sample Quantiles
0
50
100
150
-2
-1
0
1
2
Theoretical Quantiles

Area (ar = x[,8])

```
> summary(ar)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1049   36985   54277   70736   81162  566432
```

Da sich die Werte des Mittelwerts und Medians unterscheiden und beide Werte nicht ziemlich vorne in der range der Daten liegen ist eine weitere Exploration nötig. Man könnte derzeit eventuell davon ausgehen, dass die Daten in eine Richtung mehr verzerrt sind.

1) hist(ar, freq=F, breaks=5, main=paste("Area histogram of all 50 states in America"))

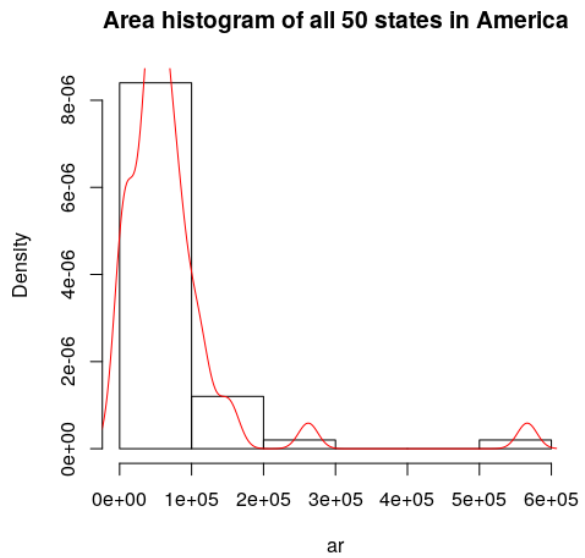
lines(density(ar), col=2)

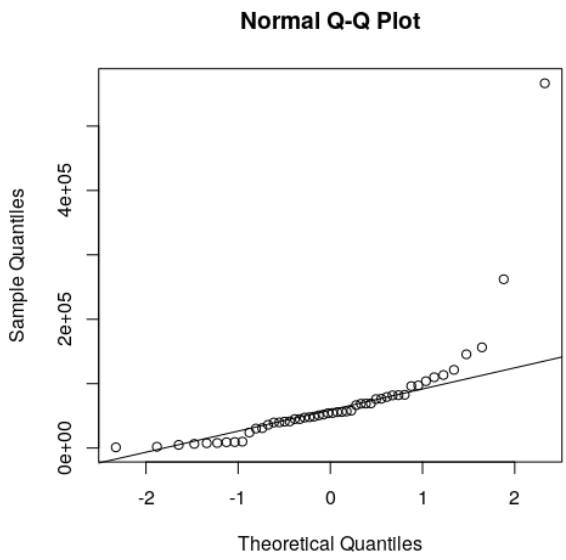
→ Histogramm zeigt, dass Daten asymmetrisch sind und die Datenpunkte vermehrt im linken Bereich der Verteilung liegen.

2) qqnorm(ar); qqline(ar)

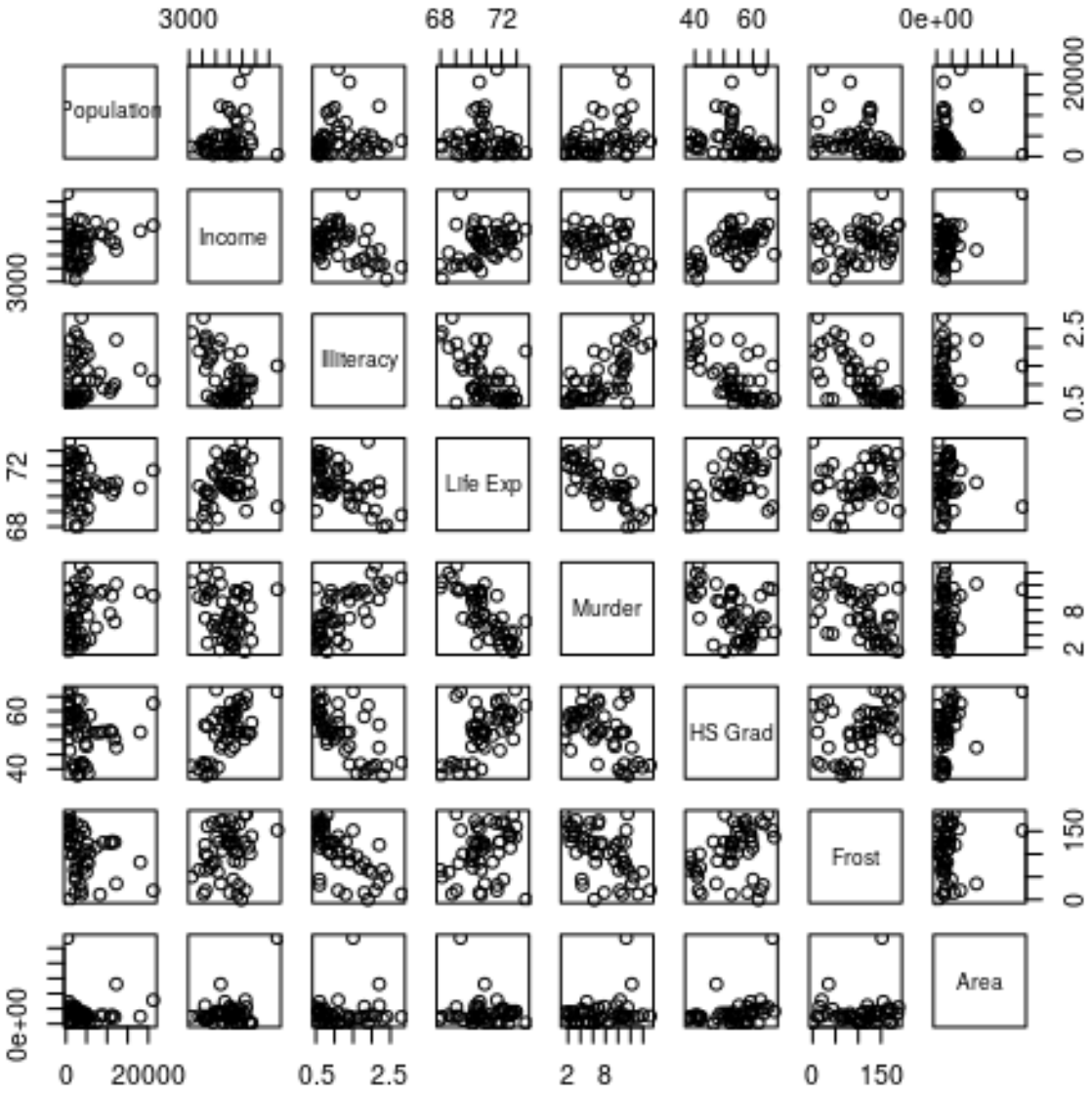
→ Anhand des qqplots sieht man, dass es einen schweren Rand (rechts vom median) gibt und einen leichten (links vom median), was auf eine rechts schiefe Verteilung hindeutet

Die Daten sind somit eindeutig nicht normalverteilt, da es keine symmetrie gibt und ein schwerer Rand vorliegt. Der Median wird in diesem Fall als Lageschätzer verwendet werden, da er der beste Lageschätzer ist, wenn die Daten nicht normalverteilt sind. Als Streuungsmaß kann man den Interquartilsabstand (IQR = 44177,25) und als Lagemaß den median absolute deviation (MAD) (mad = 35144,29) verweden.





Scatterplot



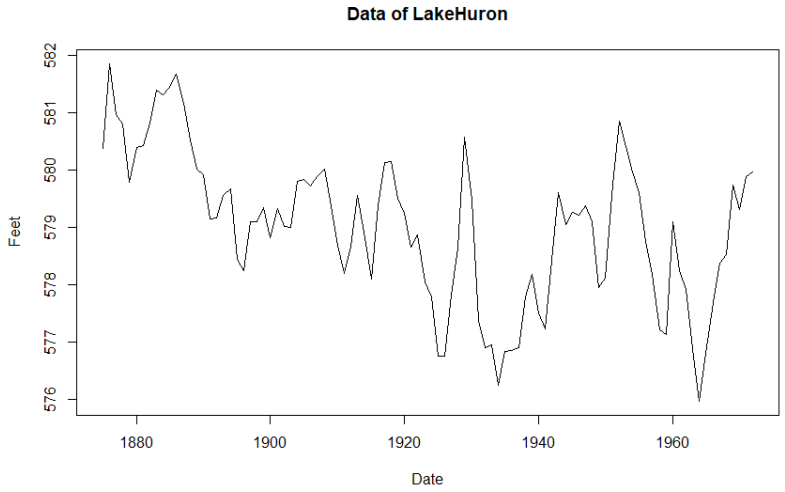

```
> cor(x)
            Population     Income  Illiteracy    Life Exp     Murder     HS Grad      Frost        Area
Population  1.00000000  0.2082276  0.10762237 -0.06805195  0.3436428 -0.09848975 -0.3321525  0.02254384
Income      0.20822756  1.0000000 -0.43707519  0.34025534 -0.2300776  0.61993232  0.2262822  0.36331544
Illiteracy  0.10762237 -0.4370752  1.00000000 -0.58847793  0.7029752 -0.65718861 -0.6719470  0.07726113
Life Exp   -0.06805195  0.3402553 -0.58847793  1.00000000 -0.7808458  0.58221620  0.2620680 -0.10733194
Murder      0.34364275 -0.2300776  0.70297520 -0.78084575  1.0000000 -0.48797102 -0.5388834  0.22839021
HS Grad    -0.09848975  0.6199323 -0.65718861  0.58221620 -0.4879710  1.00000000  0.3667797  0.33354187
Frost      -0.33215245  0.2262822 -0.67194697  0.26206801 -0.5388834  0.36677970  1.0000000  0.05922910
Area        0.02254384  0.3633154  0.07726113 -0.10733194  0.2283902  0.33354187  0.0592291  1.00000000
```

Aus dem Scatterplot sieht man, dass es eventuell eine Korrelation zwischen Income, Illiteracy, Life expectaion und Murder gibt, da sich eine leichte Linie ersichbar ist, es ist aber schwierig zu erkennen ob dies wirklich der Fall ist. Mit einer Korrelationsmatrix (nach Spearman) sieht man veremeht Werte die über 0,5 sind.

Aufgabe 3

Daten

Bei den Datensatz “LakeHuron” handelt es sich um Messdaten der jährlich gemessenen Tiefe des Lake Huron in Fuß. Der Datensatz enthält 98 Messdaten von 1875–1972. Die unveränderten Daten können mittels Plot untersucht werden.



Mittelwert:	579.0 Feet
Median:	579.1 Feet
Minimaler Wert:	575.9 Feet
Maximaler Wert:	581.8 Feet
Varianz:	1.737 Feet
Standardabw.:	1.318 Feet

Schwere Ränder und Ausreißer

Die Daten weisen einen links leicht schweren Rand auf. Es ist kein Ausreißer in den Daten (siehe QQ-Plot und Boxplot)

Schiefe

Die Daten des Histogramms zeigen nach links leicht schiefe Daten. Dies wird auch vom QQ-Plot der LakeHuron Daten bestätigt. Allerdings sind Median und Mittelwert sehr nahe beieinander. Aufgrund der nur leicht schiefen Daten sind diese dennoch normalverteilt. Die Normalverteilung wird auch beim QQ-Plot sichtbar. Es sind annähernd alle Daten auf der Linie der Normalverteilung. Lediglich die ersten Datenpunkte zeigen einen schweren linken Rand, welcher sich jedoch nach wenigen Datenpunkten wieder der Normalverteilung annähert.

Normalverteilung

Die Daten sind trotz der leichten Schiefe und des schweren linken Rands normalverteilt.

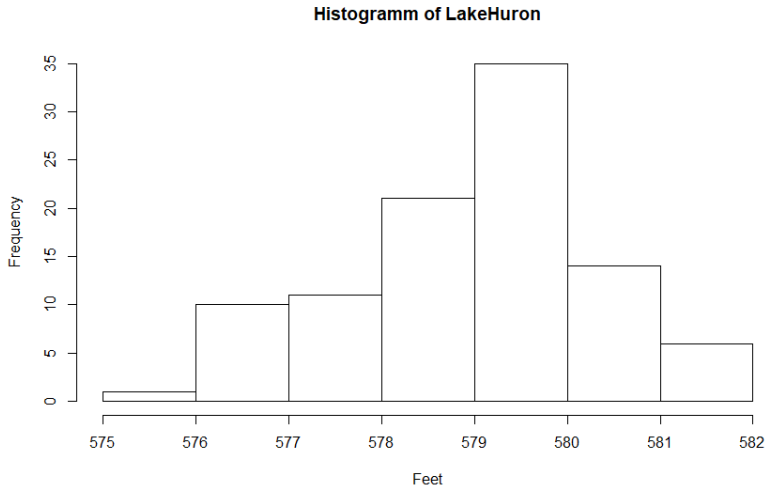
Die Vielzahl der Visualisierungen wurde nur in dieser Hausübung verwendet, um die Möglichkeiten der Datenvisualisierung aufzuzeigen. In folgende Hausübungen werden nur benötigte Visualisierungen verwendet

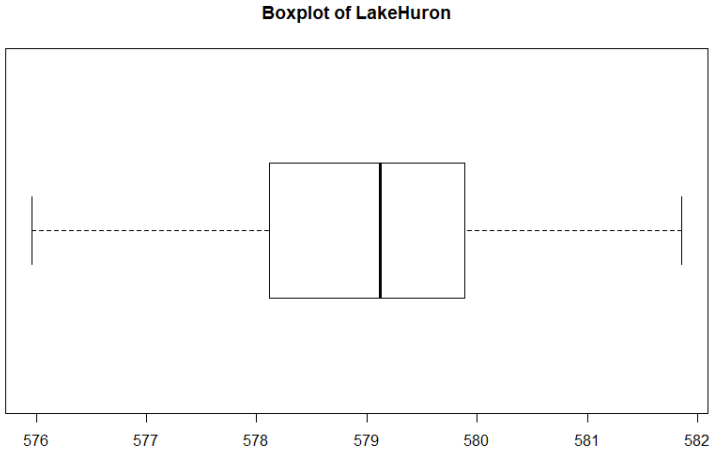
Zeitreihenaspekt

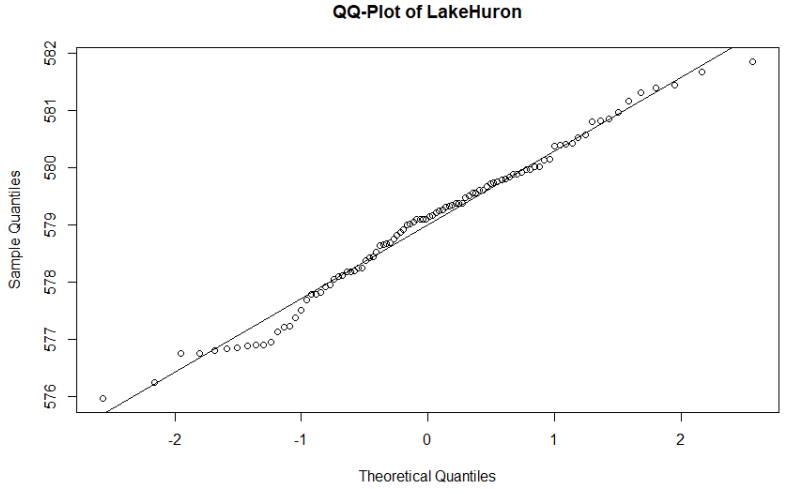
Man kann am Plot einen Trend zu einem sinkenden Wasserstand erkennen. Dies zeigt sich durch den zweiten Messwert der Datenreihe (maximaler Wert). Auch wenige Jahre später sind sehr hohe

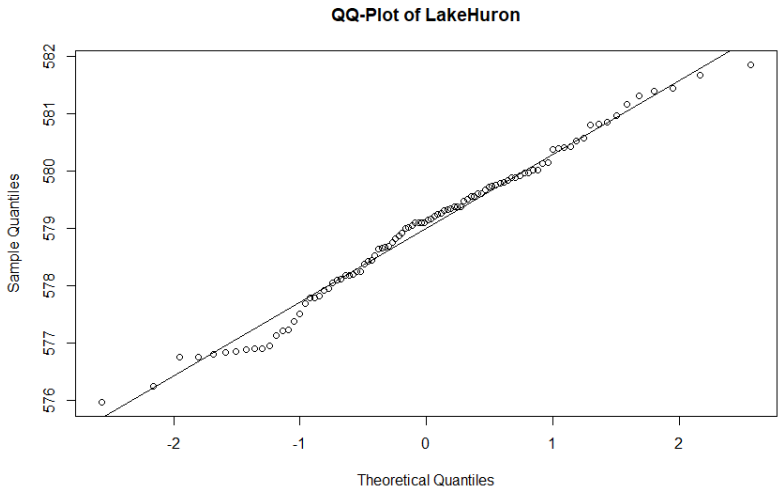
Wasserstände verzeichnet worden. Der Minimalwert wurde hingegen nach 1960 gemessen. Auch die Jahre zwischen 1920 und 1940 weißen sehr viele niedrige Wasserstände auf.

Visualisierungen










```
#####
# Schätzer
#####
cat("Mean:         ", mean(x), "\n")       # Durchschnitt
cat("Median:       ", median(x), "\n")     # Median
cat("Min:          ", min(x), "\n")        # Minimaler Wert
cat("Max:          ", max(x), "\n")        # Maximaler Wert
cat("Variance:     ", var(x), "\n")        # Varianz
cat("Standardev.: ", sd(x), "\n")          # Standardabweichung


#####
# Visualisierungen
#####
# Visualisierung der unveränderten Daten
plot(x, main="Data of LakeHuron", ylab="Feet", xlab="Date")
# Boxplot der Messdaten
boxplot(x, main="Boxplot of LakeHuron", horizontal = TRUE)
# Histogramm der Messdaten
hist(x, xlab="Feet", main="Histogramm of LakeHuron")
# QQ-Plot der Messdaten
qqnorm(x, main="QQ-Plot of LakeHuron"); qqline(x)
```

Aufgabe 4

Der Datensatz Titanic enthält Informationen bezüglich des Schicksals der Passagiere, ihres Alters, Geschlechts sowie ihrer sozioökonomischen Klasse. In tabellarischer Form seien die genauen Zahlen dargestellt:

			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

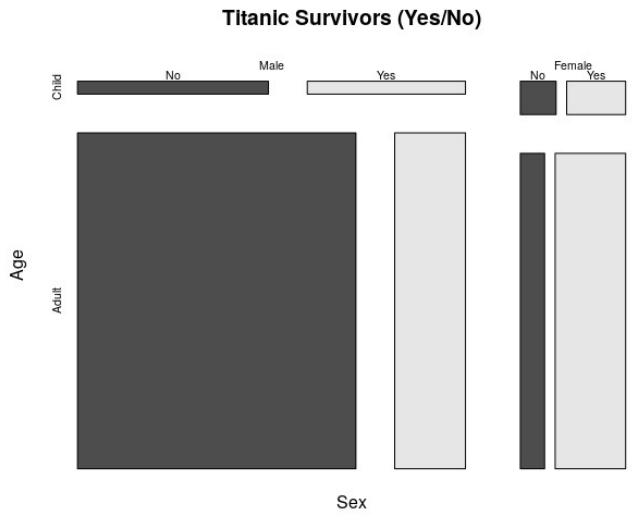
Von den insgesamt 1731 männlichen Passagieren sind 1364 verstorben (21,20% Überlebensrate), von den 470 weiblichen Passagieren sind 126 verstorben (73,19% Überlebensrate).

Weiters sind von den 109 Kindern 52 verstorben bei einer Überlebensrate von 52,29%.

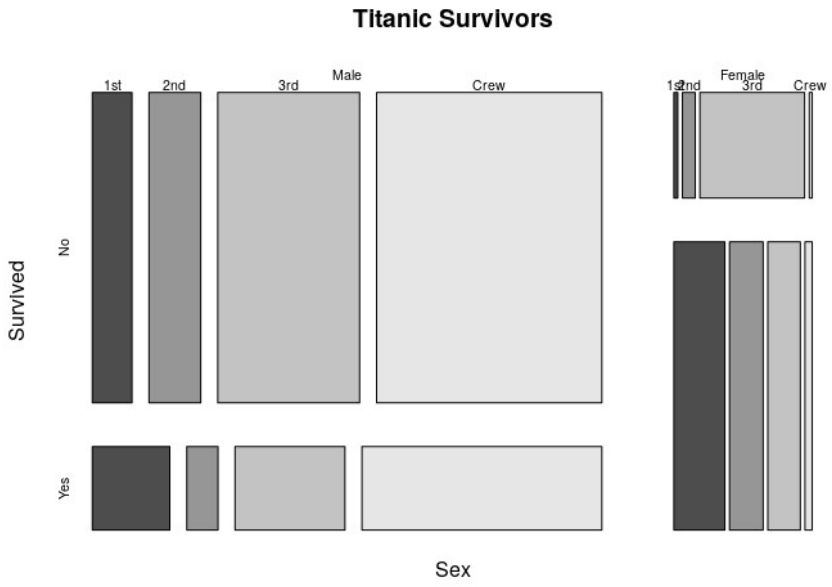
Von den Klassen ergeben sich folgende Überlebensraten:

	Survived		
	No	Yes	c_prop
1st	122	203	0.6246
2nd	167	118	0.4140
3rd	528	178	0.2521
Crew	673	212	0.2395

mit `“c_prop”` als Spalte für die Überlebensraten. Hier ist ein deutlicher Unterschied zwischen den Klassen zu erkennen, von Klasse 1 mit über 60% verglichen mit Klasse 3 und den Crewmembern mit rund 25%.
Mittels Mosaicplots lassen sich die Daten visuell übersichtlich darstellen:



Auf ersten Blick fallen die wesentlich größeren Balken der verstorbenen Männerfraktion auf, während der Zahl der Überlebenden zwischen beiden Geschlechtern insgesamt etwa gleich groß ist (367 Männer gegen 344 Frauen).



Betrachtet man die Klassenaufteilung im nächsten Mosaikplot ergibt sich dasselbe Bild wie auch oben bei der Tabelle: Die Passagiere niedrigerer Klassen oder Schiffsmitglieder hatten eine geringere Chance zu überleben als die der höheren.

Simpsons Paradoxon

Wenn man z.B. die Überlebensrate der dritten Klasse im Vergleich zur Crew vergleicht, dann hat die dritte Klasse eine höhere Überlebenschance. Wenn man aber die Daten in Geschlechter aufteilt sieht man, dass die Crew eine bessere Überlebenschance hat (bei Frauen sowie bei Männern). Die erste Annahme wurde somit durch eine weitere Aufteilung der Daten verändert.

Annahme: Es gibt mehr Männer in der Crew als in der dritten Klasse und mehr Frauen in der dritten Klasse als in der Crew. Da Männer im Vergleich zu Frauen eine geringere Überlebensrate hatten und überwiegend Männer in der Crew waren könnte man annehmen, dass die Crew eine schlechtere Überlebenschance hat.

→ Erst nach genauerem Betrachten der Daten für Frauen und Männer sieht man, dass dies nicht der Fall ist.