Praktikums Bericht

Tobias Chisi

Tag .1

Aufgabe 1

a) Mit diesem Befehl wurde der Datei 0ecdM.csv heruntergeladen.

```
OECD_Data <- read.csv("Data/oecdM.csv", header = TRUE, sep= ",",
dec= ".", stringsAsFactors =  FALSE)
```

b) Von der Daten Satz haben wir ein Mittelwert von

- Einkommen = 19.183
- Armut = 12.37667
- Bildung = 2.673333
- WenigRaum = 31.95385
- Umwelt = 25.22083
- Lesen = 496.32
- Geburtsgewicht = 6.43
- Säuglsterblichkeit = 5.446667
- Sterblichkeit = 24.6069
- Selbstmord = 6.851724
- Bewegung = 20.13462
- Rauchen = 16.5125
- Alkohol = 15.225
- Jugendschwanger = 15.5
- Bullying = 10.97917
- Schule = 27.172

Und eine Varianz von

- Einkommen = 50.75937
- Armut = 31.32599
- Bildung = 10.99168
- WenigRaum = 446.9546
- Umwelt = 56.1052
- Lesen = 862.9761
- Geburtsgewicht = 3.708379
- Säuglsterblichkeit = 20.31085
- Sterblichkeit = 45.51852
- Selbstmord = 10.26116
- Bewegung = 37.92075
- Rauchen = 22.72288

- Alkohol = 18.50543
- Jugendschwanger = 195.0862
- Bullying = 26.48607
- Schule = 108.1446

c) Niederlande ist in der Daten Satz und China nicht.

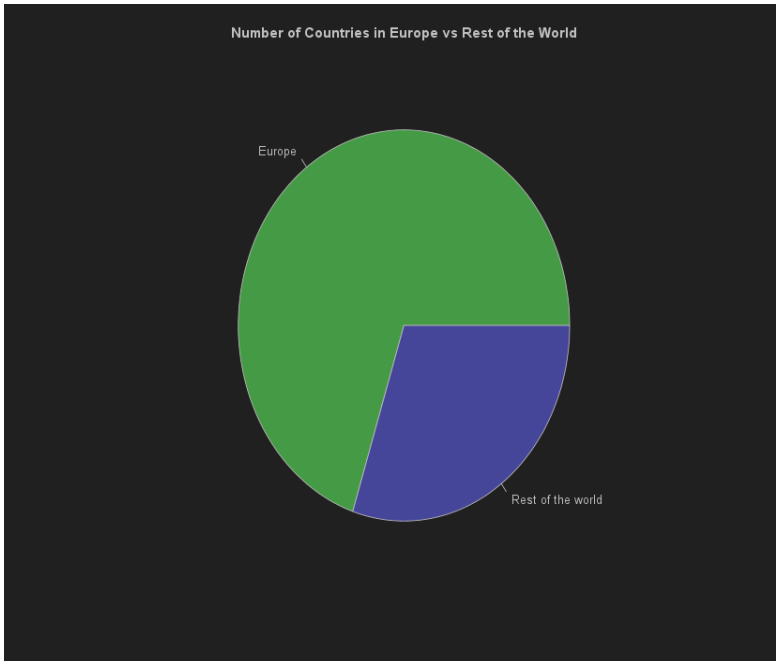
d) In die Länder: Österreich, Belgien, Kanada, Tschechien, Dänemark, Finnland, Frankreich, Deutschland, Griechenland, Ungarn, Island, Irland, Italien, Luxemburg, Niederlande, Norwegen, Polen, Portugal, Slowakei, Spanien, Schweden, Schweiz, Vereinigtes Königreich, USA
waren die meisten Jugendlichen mindestens zweimal betrunken. Die mit dem höchsten maximalen Prozentsatz war Dänemark.

e) Am geringsten ist die Säuglingssterblichkeit in Island mit 2,3

f) Der durchschnittliche Prozentsatz von Jugendlich die sich regelmäßig bewegen ist 20.13463 und die Länder mit eine kleiner durschnitt sind: Österreich, Belgien, Frankreich, Deutschland, Griechenland, Ungarn, Italien, Luxemburg, Mexiko, Norwegen, Polen, Portugal, Schweden, Schweiz, Türkei, Vereinigtes Königreich

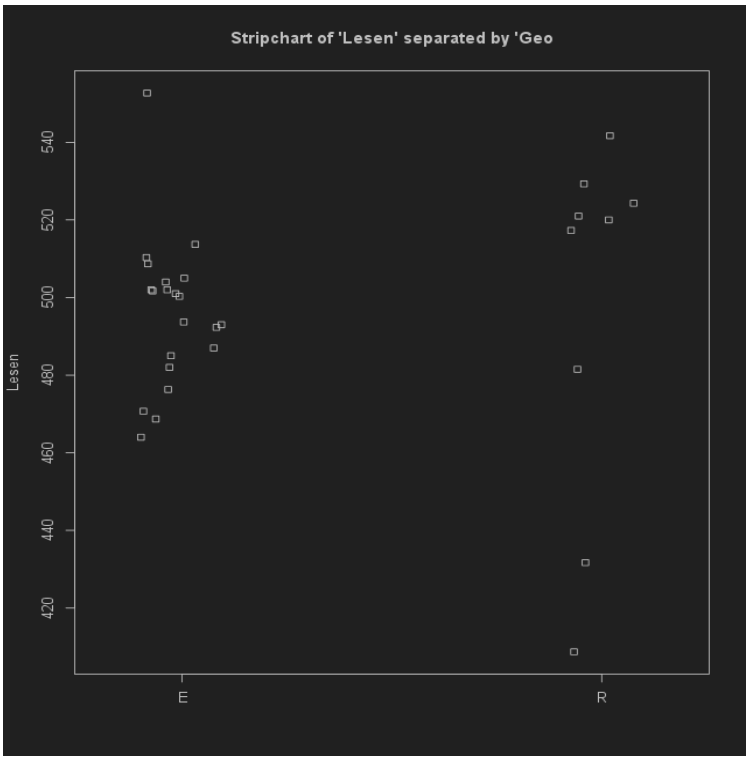
Aufgabe 2

a) Das sind 21 Länder aus dem Europa und 9 aus dem Rest der Welt.



Das seht man auch hier im Kuchendiagramm

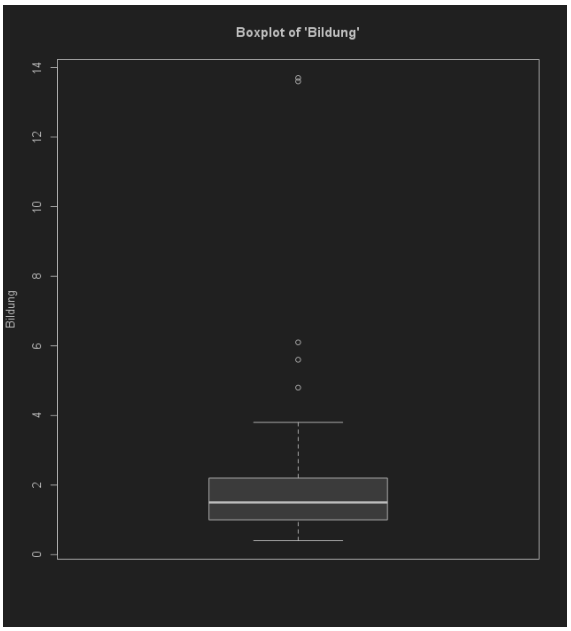
b)



Von der Stripchart können wir die Aussage treffen das europäische Länder mindesten ein score von 460 beim Lesen haben. Der Rest der Welt variiert viel mehr mit ihr Lesen score in verglich zu Europa, wo die Lese score zwischen 460 und 520.

Aufgabe 3

a)



Q1 und Q2 sind sehr nah beieinander, also sehr nah an dem Mittelwert. Q3 hat eine größere flache als Q1 und wir haben eine paar Ausreißer nahe bei 14.

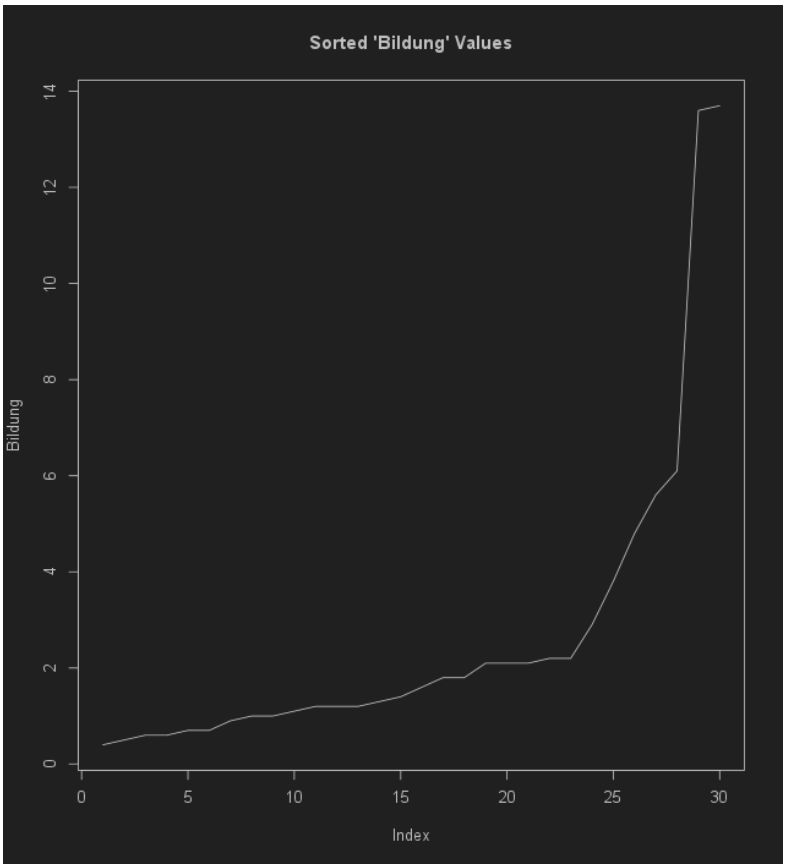
b)

Nach Berechnung mit quantile() kriegen wir

0%	25%	50%	75%	100%
0.4	1.0	1.5	2.2	13.7

Hier sehe wir das ungefähr 50% die daten zwischen 1 und 2.2 legen und die flache zwischen 0% - 25% kleiner ist als 50% - 75%

c)



Der Graph ist exponentiell

d)

Zwischen 75% und 100% haben wir ein Sprung von 2.2 bis 13.7 das ist einen großen Unterschied und von 0% bis 75% haben wir ein Sprung von 0.4 bis 2.2 das ist minimal in vergleich. Das ist auch zu sehen in der obenstehend Line Graph ab 2.2 geht der Line sehr schnell hoch. Also ist ab 75% ein exponentielles Wachstum. Dadurch können wir sagen das 75% eine guten Trennpunkt zwischen Ländern mit “guter” und “schlechter” Grundausstattung für Bildung darstellt.

Tag .2

Aufgabe 4

a)

```
X1 <- rexp(n = 100, rate = 0.1)
HX2 <- rexp(n = 100, rate = 0.1)
X2 <- 20 - HX2
```

b)

Welch Two Sample t-test

data: X1 and X2

t = -0.043913, df = 194.81, p-value = 0.965

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -3.214811 3.074768

sample estimates:

mean of x mean of y

 10.63709 10.70711

Die t wert ist bei -0.043913 also sind beide Median werte sehr ähnlich. Da der p wert fast bei 1 ist können wir die 0 Hypothese nicht ablehnen.

c)

Wilcoxon rank sum test with continuity correction

data: X1 and X2

W = 3580, p-value = 0.0005236

alternative hypothesis: true location shift is not equal to 0

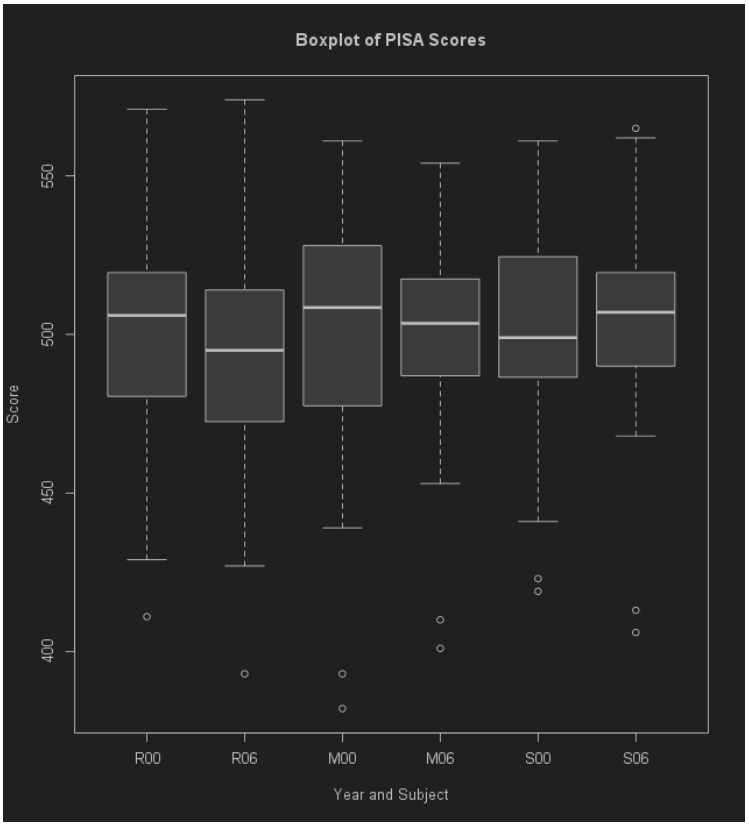
Die p-wert ist bei 0.0005236 also sehr niedrig, das heißt wir können die 0 Hypothese ablehnen und die Verteilungen sind nicht gleich.

Aufgabe 5

a)

```
PISA_Data <- read.csv("Data/PISA.csv", header = TRUE, sep= ",", dec= ".",
stringsAsFactors =  FALSE)
```

b)



Bei Lesekompetenz sehen wir einen verschlechterten Median von 2000 zu 2006. Das ist genau so der Fall bei Mathematik aber nicht Naturwissenschaften. Von 2000 zu 2006 hat sich der Median von Mathematik verschlechtert aber den Abstand zwischen Q1 und Q3 sind kleiner geworden.

c)

Paired t-test

data: PISA_Data$M00 and PISA_Data$M06

t = -0.008081, df = 51, p-value = 0.9936

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

-4.796802 4.758340

sample estimates:

mean difference

-0.01923077

Da der p-wert bei 0.9 legt könne wir die 0 Hypothese nicht ablehnen. Das heißt dass es keine signifikante Änderung gibt zwischen M00 und M06.

Paired t-test

data: PISA_Data$R00 and PISA_Data$R06

t = 2.2964, df = 51, p-value = 0.0258

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

0.7159095 10.6687059

sample estimates:

mean difference

5.692308

Da der p-Wert bei 0.02 und kleiner 0.05 können wir die 0 Hypothese ablehnen also es gibt eine signifikante Veränderung zwischen die PISA-Scores von R00 und R06.

Paired t-test

data: PISA_Data$S00 and PISA_Data$S06

t = -1.2842, df = 51, p-value = 0.2049

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

 -7.443524 1.635832

sample estimates:

mean difference

     -2.903846

Da der p-Wert bei 0.2 liegt können wir die 0 Hypothese nicht ablehnen. Das heißt dass es keine signifikante Änderung gibt zwischen M00 und M06

Aufgabe 06

One Sample t-test

data: Hustensaft_Data$Kon

t = -1.7586, df = 8, p-value = 0.1167

alternative hypothesis: true mean is not equal to 40

95 percent confidence interval:

 38.15097 40.24903

sample estimates:

mean of x

    39.2

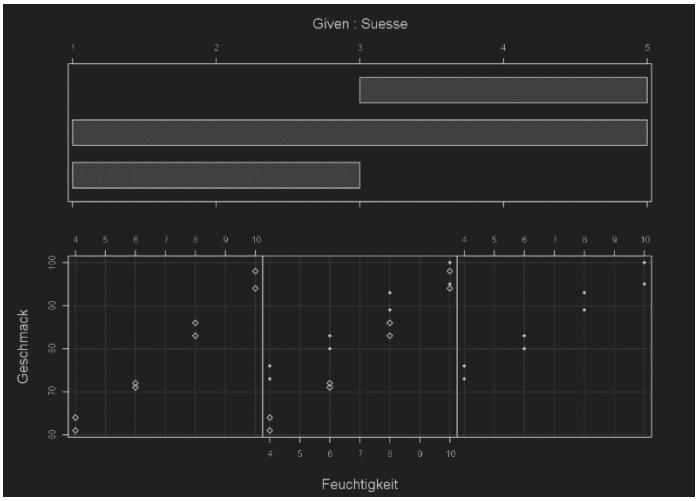
Da der p-value höher als 0.05 ist können wir die 0 Hypothese nicht ablehnen. Die Stichprobe hatte eine median von 39.2 also auch nicht so eine Größe unterschied. Also haben wir keine Begründung für eine Produktionstop.

Aufgabe 07

a)

```
sues <- read.csv("Data/Suess.csv", header = TRUE, sep = ";", dec = ".",
stringsAsFactors = FALSE)
```

b)



Anhand der obenstehenden Grafik können wir sehen das umso süßer umso hoher ist der Geschmack und umso mehr Feuchtigkeit vorhanden ist. Bei einer Süßlichkeit von 2 habe wir eine niedriger mindestwert als bei Süßlichkeit 4. Der Feuchtigkeit führt zu einer liniere Steigerung der Geschmack bei 2 und 4.

C)

Call:

lm(formula = Geschmack ~ Feuchtigkeit + Suesse, data = sues)

Residuals:

Min	1Q	Median	3Q	Max
-5.525	-1.850	-0.325	1.775	4.975

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	37.5250	3.3451	11.218	4.67e-08	***
Feuchtigkeit	4.8000	0.3362	14.277	2.53e-09	***
Suesse	3.7500	0.7518	4.988	0.000248	***

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

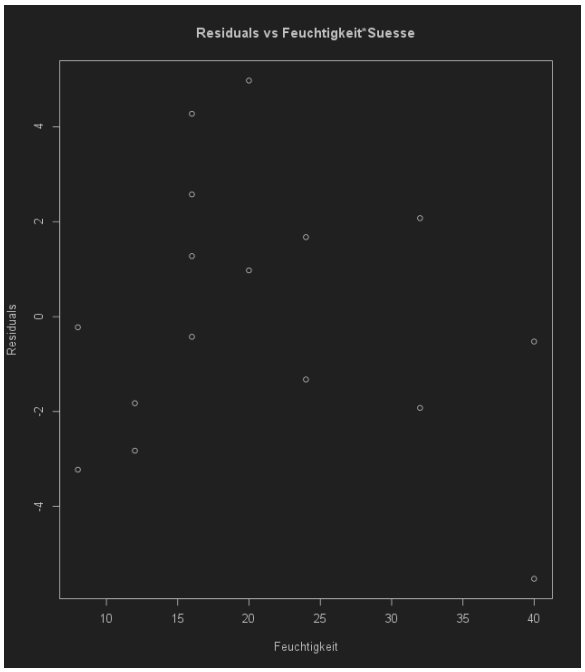
Residual standard error: 3.007 on 13 degrees of freedom

Multiple R-squared: 0.9462, Adjusted R-squared: 0.9379

F-statistic: 114.4 on 2 and 13 DF, p-value:

Hier wurde ein lineares Modell mit Geschmack als abhängige Variable und Feuchtigkeit und Suesse als unabhängige Variablen erstellt. Die p-wert ist 5.611e-09 d.h. mindestens eine unabhängige Variable hat eine signifikante Beziehung zur abhängigen Variable.

d)



Aus den obigen Daten können wir erkennen, dass es einen Aufwärtstrend gibt. Je höher die Feuchtigkeit*Suesse, desto höher der Residual, auch wenn wir einige Ausreißer haben.

e)

Call:

lm(formula = Geschmack ~ Feuchtigkeit + Suesse + Feuchtigkeit:Suesse,

   data = sues)

Residuals:

  Min   1Q Median   3Q   Max

-2.900 -1.425 -0.775 1.800 2.950

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept)   19.1500   5.6275   3.403 0.005241 **

Feuchtigkeit 7.4250   0.7658   9.696   5e-07 ***

Suesse 9.8750 1.7796 5.549 0.000126 ***

Feuchtigkeit:Suesse -0.8750 0.2422 -3.613 0.003559 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

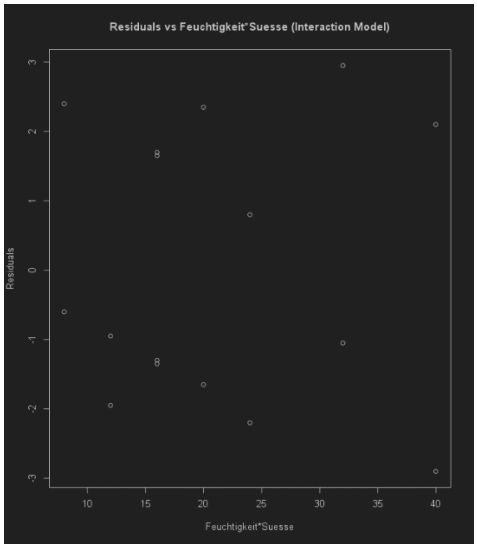
Residual standard error: 2.166 on 12 degrees of freedom

Multiple R-squared: 0.9742, Adjusted R-squared: 0.9678

F-statistic: 151.3 on 3 and 12 DF, p-value: 8.469e-10

sowohl Feuchtigkeit als auch Suesse hat einen signifikanten Einfluss auf Geschmack.

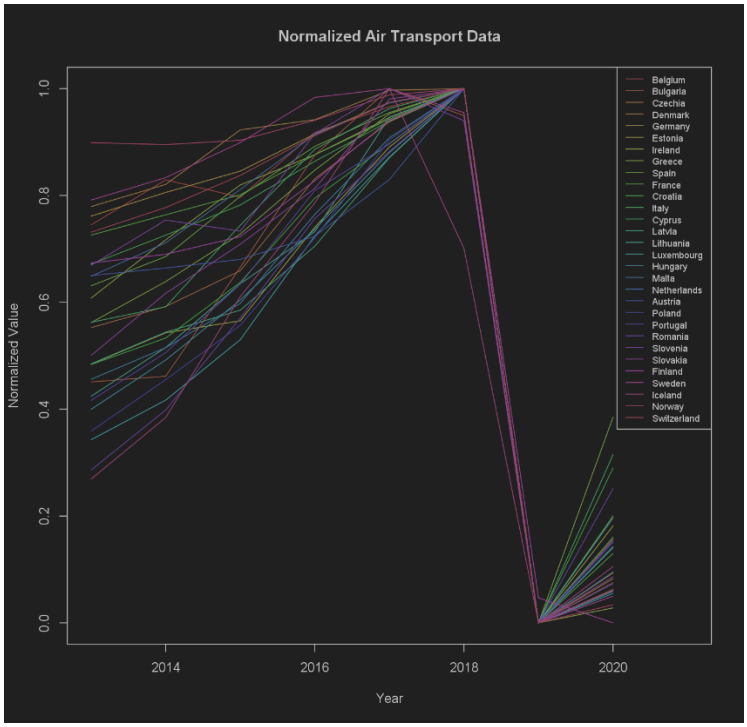
f)



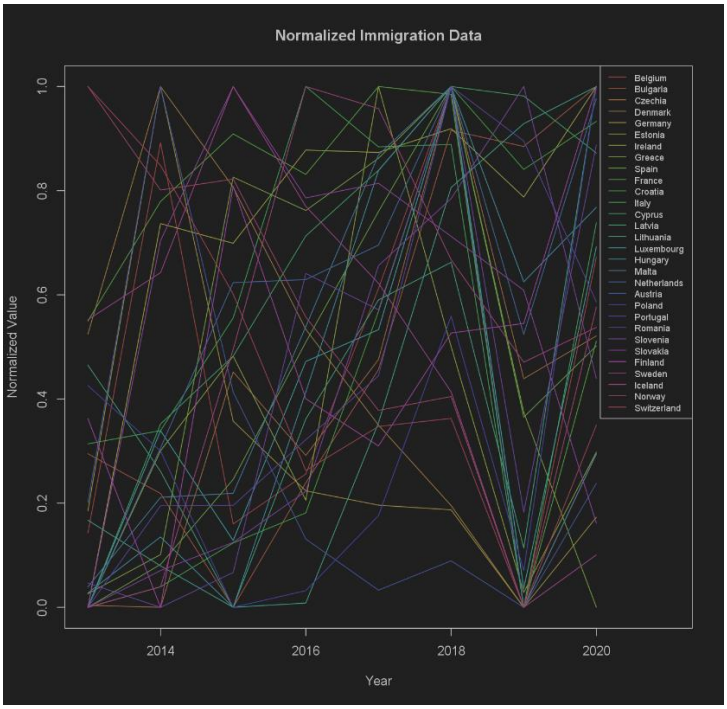
Anhand der obigen Daten können wir sehen, dass es einen Abwärtstrend gibt. Je höher die Feuchtigkeit*Suesse, desto niedriger der Residual, auch wenn wir einige Ausreißer haben.

Aufgabe 08

a)

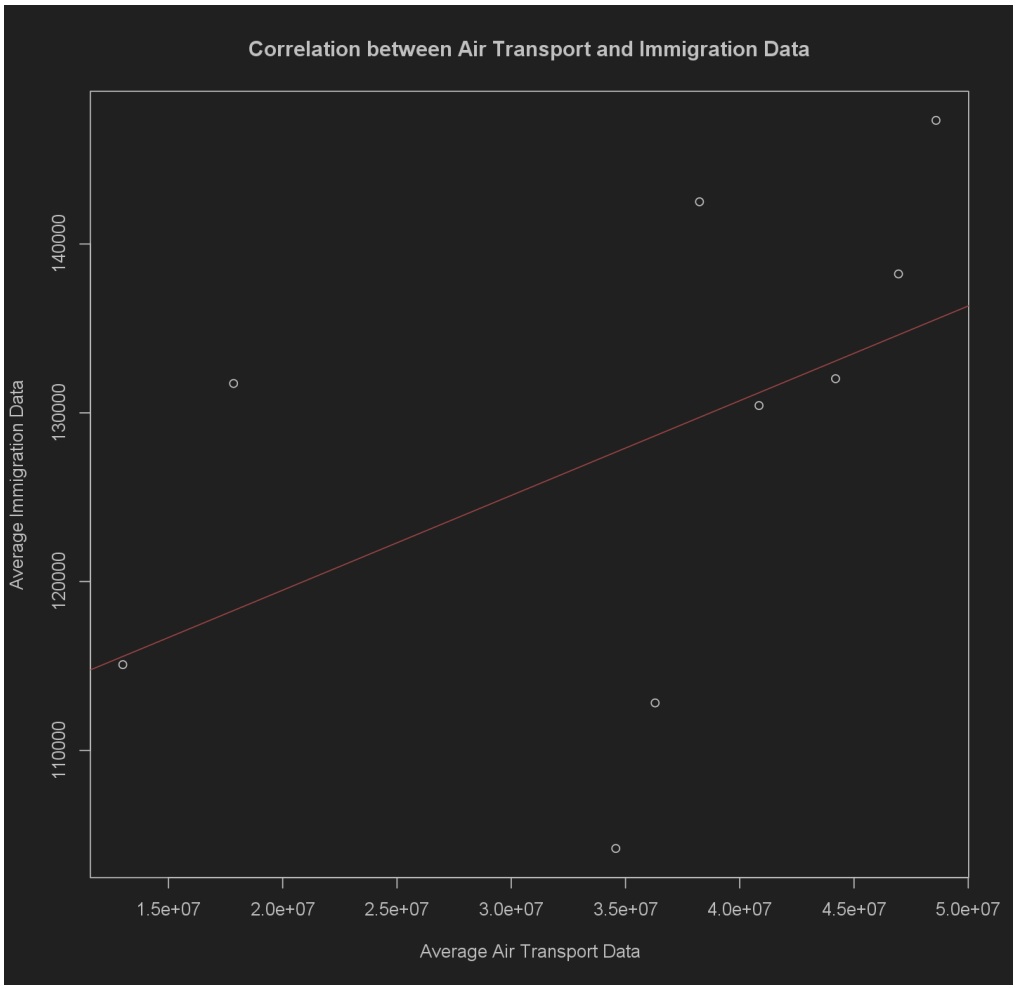


Die obige Grafik verdeutlicht deutlich eine drastische Senkung des Flugverkehrs im Jahr 2019. Diese Entwicklung lässt sich klar auf die Auswirkungen des Coronavirus und der damit verbundenen Lockdown-Maßnahmen zurückführen.



In der zweiten Grafik (Normalized Immigration Data) können wir erkennen, dass die Daten deutlich stärker zufällig verteilt sind, obwohl wir auch in einigen Ländern einen klaren Tiefpunkt im Jahr 2019 sehen.

b)



Die Korrelation zwischen den Durchschnittswerten der Air Transport Data und den Immigration Data beträgt 0,479665908858791. Der p-Wert für die Korrelation beträgt 0,191333682344951. Die Korrelation von 0,479 zwischen den Durchschnittswerten der Air Transport Data und den Immigration Data deutet darauf hin, dass es eine moderate positive lineare Beziehung zwischen diesen beiden Variablen gibt. Mit anderen Worten, wenn der Luftverkehr zunimmt, steigt auch die Einwanderung tendenziell und umgekehrt, obwohl diese Beziehung nicht sehr stark ist.

Der p-Wert von 0,191 deutet darauf hin, dass diese Korrelation statistisch nicht signifikant ist, da der p-Wert größer als das übliche Signifikanzniveau von 0,05 ist. Das bedeutet, dass die beobachtete Korrelation durch Zufall (bzw. Korona Pandemie) zustande gekommen sein könnte und nicht auf eine tatsächliche Beziehung zwischen den Variablen hinweist.

c)

Für den Datensatz Air Transport weist die Schweiz die höchste Korrelation mit Deutschland auf, während Rumänien die niedrigste Korrelation aufweist.

Für den Datensatz Immigration weist Österreich die höchste Korrelation mit Deutschland auf, während Finnland die niedrigste Korrelation aufweist.

Tag 3

Aufgabe 9

a)

```
load("Data/SPECTF.RData")
objects <- ls()
SPECTF_Dimensions <- dim(SPECTF)
```

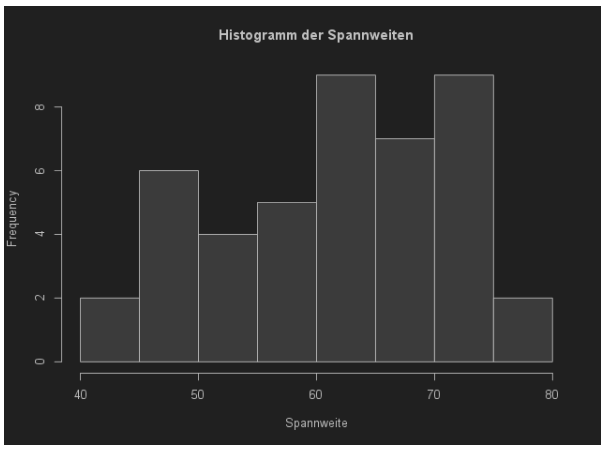
objects : "objects" "SPECTF" "SPECTF_Dimensions"

SECTF Dimension: 267 45

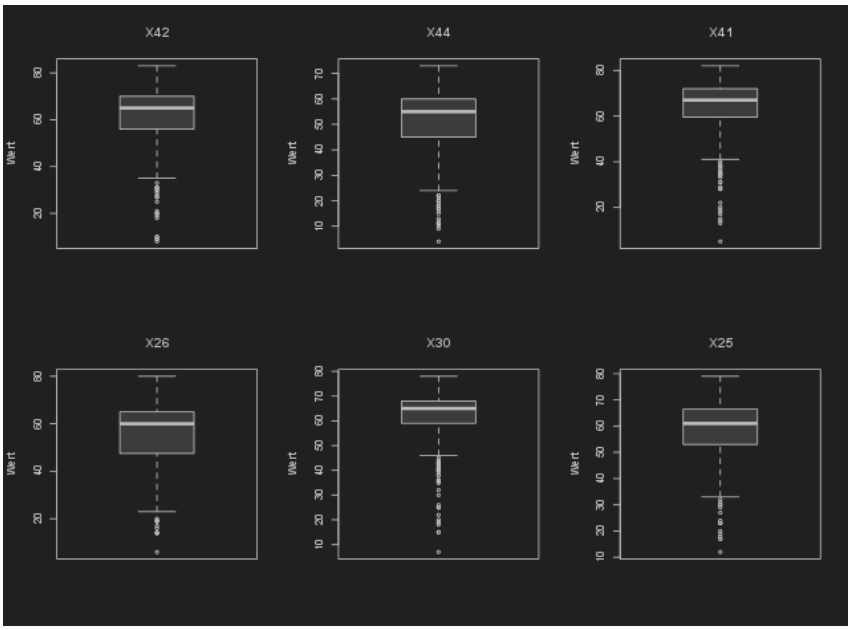
b)

Spannweite

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
41.00	53.75	64.00	62.18	70.25	77.00



c)



Aufgabe 10

a)

Kreuzvalidierungsverfahren sind statistische Methoden zur Bewertung und Optimierung von Modellen in der maschinellen Datenanalyse. Eine häufig verwendete Methode ist die K-Fold Cross-Validation, bei der der Datensatz in K Teile aufgeteilt wird. Das Modell wird iterativ auf K-1 Teilen trainiert und auf dem verbleibenden Teil getestet. Dieser Prozess wird wiederholt, und die Leistung des Modells wird durchschnittlich bewertet. Kreuzvalidierung ist wichtig, um die Robustheit eines Modells zu gewährleisten und Überanpassung an die Trainingsdaten zu vermeiden. Es ist eine gängige Praxis in der Modellbewertung und -auswahl.

d)