1)a) Zuerst muss man die Datei oecdM.csv von Ilias hochladen. Danach muss man die Datei mit read.table() Befehl lesen. Man nutzt auch attach-Befehl um einfacher mit den Daten umgehen zu können.

b)Für diese Aufgabe nutzt man sapply(X,FUN,na.rm=TRUE), weil es Matrix ist. X ist Matrix. FUN ist Funktion. na.rm=TRUE ausschließt fehlender Wert, wenn es deskriptive Statistik in R kalkuliert.
Mittelwert:

Einkommen	Armut	Bildung	WenigRaum
19.183333	12.376667	2.673333	31.953846
Umwelt	Lesen	Geburtsgewicht	Säuglsterblichkeit
25.220833	496.320000	6.430000	5.446667
Sterblichkeit	Selbstmord	Bewegung	Rauchen
24.606897	6.851724	20.134615	16.512500
Alkohol	Jugendschwanger	Bullying	Schule
15.225000	15.500000	10.979167	27.172000

Varianz:

Einkommen	Armut	Bildung	WenigRaum
50.759368	31.325989	10.991678	446.954585
Umwelt	Lesen	Geburtsgewicht	Säuglsterblichkeit
56.105199	862.976138	3.708379	20.310851
Sterblichkeit	Selbstmord	Bewegung	Rauchen
45.518522	10.261158	37.920754	22.722880
Alkohol	Jugendschwanger	Bullying	Schule
18.505435	195.086207	26.486069	108.144600

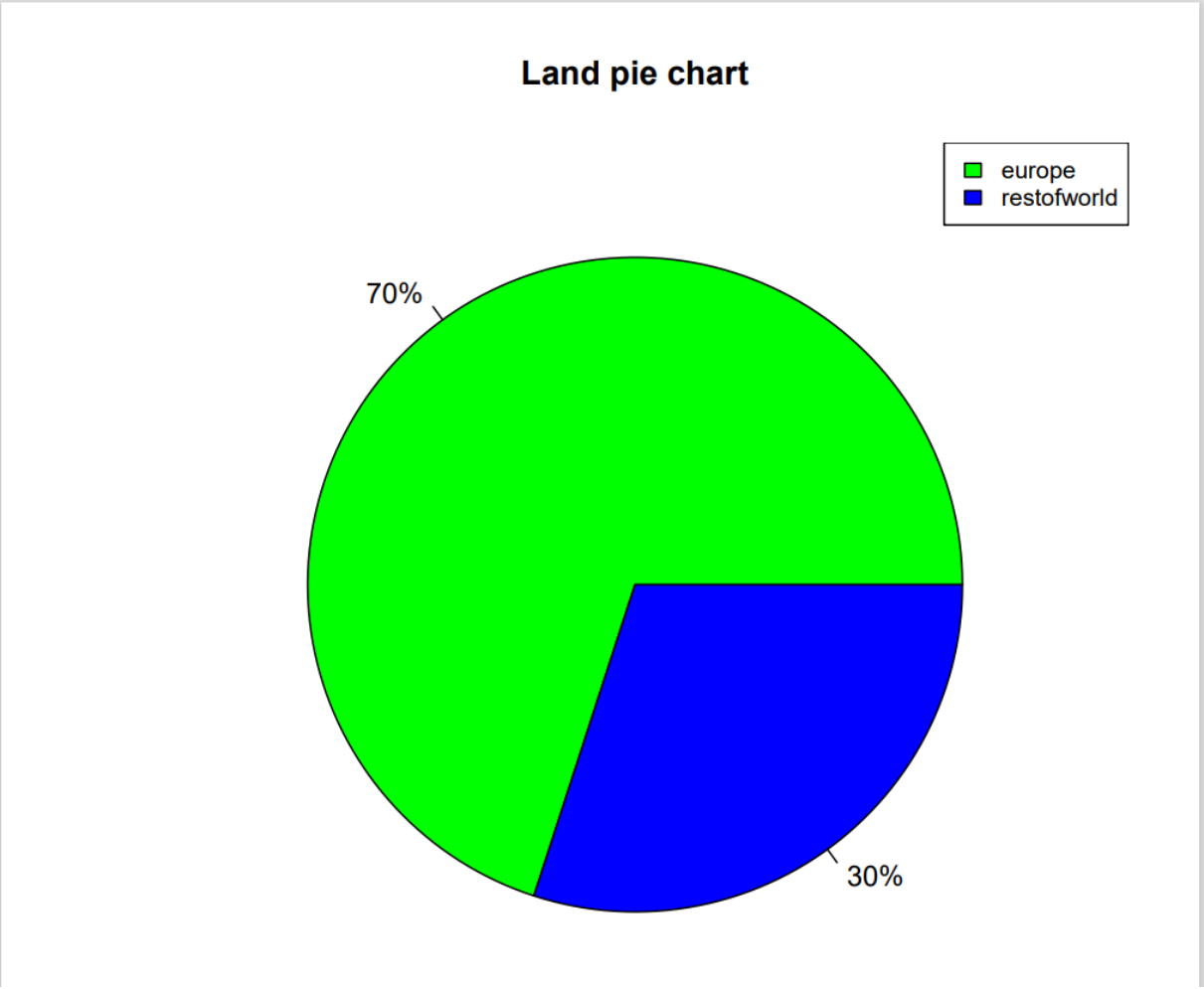
c)Man nutzt grepl(“pattern“,x) Befehl in R. grepl (grep logical) ist eine Funktion, die richtig, wenn “pattern“ in x gibt. “pattern“ ist die Eintrag. x ist die subset von daten. Es gibt Niederlande in der Länderliste des Datensatze. Es gibt keine China in der Länderliste des Datensatze.

d)Für die maximale Prozentsatz nutzt man max(x,na.rm=TRUE). x ist die subset von daten. na.rm=TRUE ausschließt fehlender Wert, wenn es deskriptive Statistik in R kalkuliert. Maximale Prozentsatz ist 24.8. Für das Land mit den meisten Jugendlichen mindestens zweimal betrunken ist Dänemark. Man nutzt while Befehl in R.

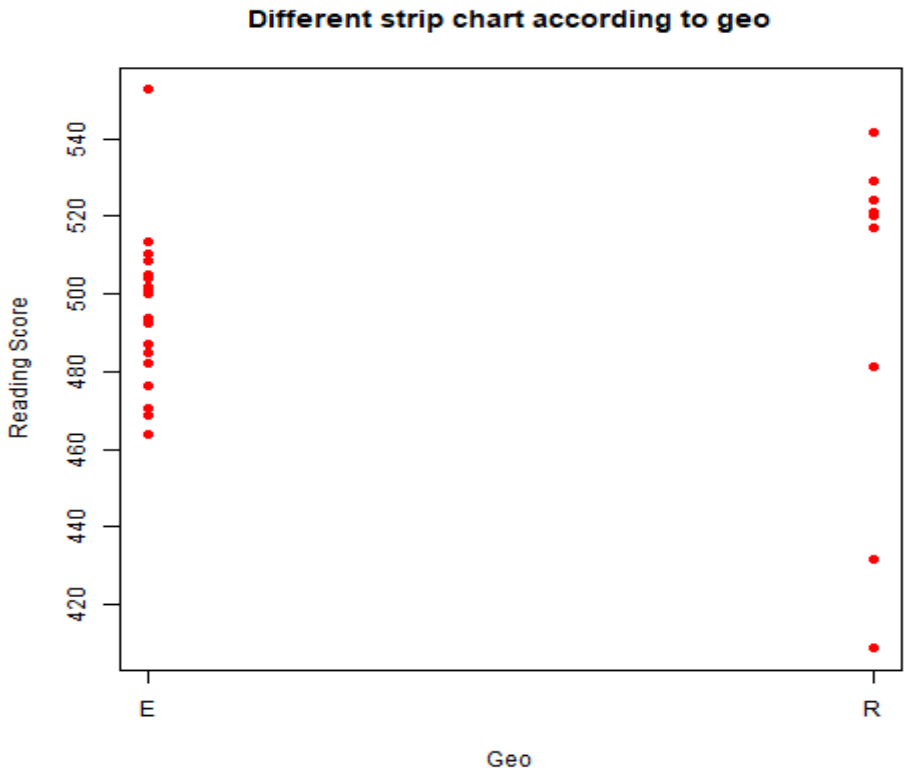
e) Für die minimale Prozentsatz nutzt man min(x,na.rm=TRUE). x ist die subset von daten. na.rm=TRUE ausschließt fehlender Wert, wenn es deskriptive Statistik in R kalkuliert. Minimale Prozentsatz ist 2.3. Island ist das Land mit dem geringsten Säuglsterblichkeit. Man nutzt while Befehl in R.

f)Man nutzt while Befehl in R. Länder mit der Prozentsatz an Jugendlichen, die sich regelmäßig bewegen, kleiner als durchschnitt : Es gibt 16 Länder Österreich, Belgien, Frankreich, Deutschland, Griechenland, Ungarn, Italien, Luxemburg, Mexiko, Norwegen, Polen, Portugal, Schweden, Schweiz, Türkei, Vereinigtes Königreich.

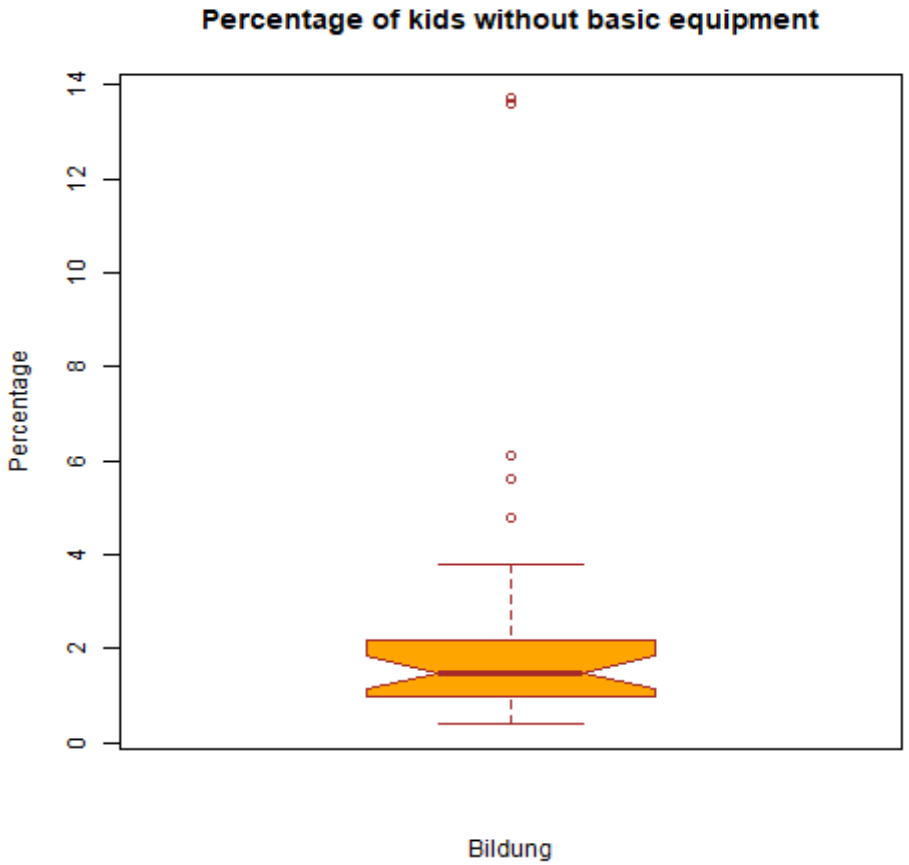
2)a)21 Länder gehören zu Europa und 9 Länder sind Rest der Welt. Für dieses Aufgabe nutzt man length(which()) Befehl in R. length ist die Länge von einem subset. Für die Küchendiagramm nutzt man mit pie(x,labels,main,col) Befehl in R. Man sieht, dass 70% von der Land in Europa liegen. Die andere 30% von Land liegen rest der Welt.



b) Für die Stripchart Lesen: Das niedrigste Lesen Punkte ist nicht aus Europa. Das höchste Lesen Punkte ist aus Europa. Das Lesen Punkte außer Europa sind so miteinander unterschiedlich.

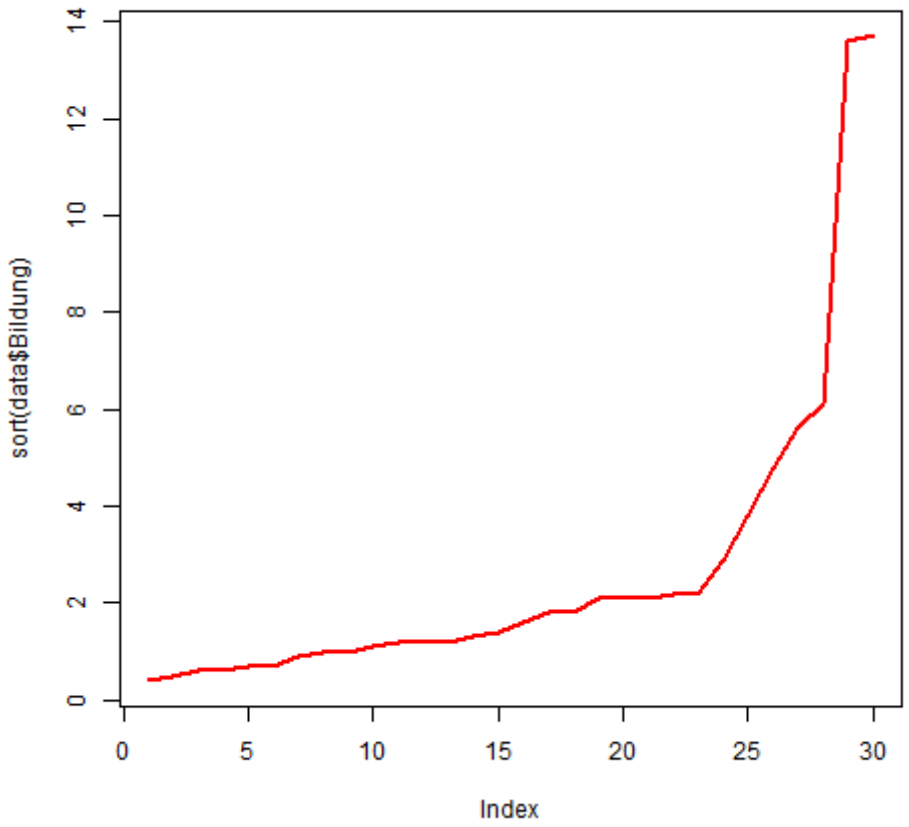


3)a) Türkei und Mexiko sind die zwei Ausreißer.



b)Man nutzt quantile() Befehl in R.
0% 25% 50% 75% 100%
0.4 1.0   1.5   2.2   13.7
Das Ergebnis in 3b und das Bild in 3a übereinstimmen.

c) Man kann sieht, dass die Datei schon von aufsteigende Werte ordnet. Je größer die Index, desto größer die Anteil der Kinder, die ohne Grundausstattung für Bildung auskommen.



d) Es ist einen guten Trennpunkt zwischen Ländern mit "guter" und "schlechter" Grundausstattung für Bildung, da für Werte, die größer als das 75% Quantil sind, steigt die Kurve offensichtlich stark an.

4)a) Für dieses Aufgabe nutzt man rexp(N,lambda) Befehl in R, weil es exponential verteilte Zufallszahlen.

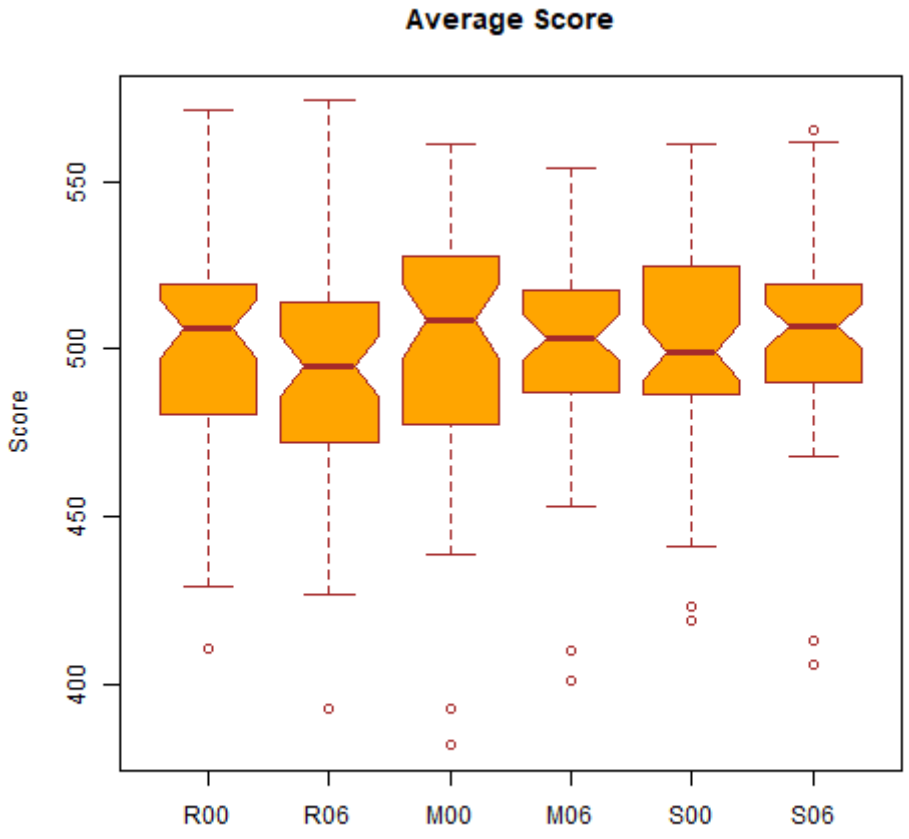
b) Welch Two Sample t-test
data: X1 and X2
t = 0.138, df = 197.71, p-value = 0.8904
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.674700 3.077225
sample estimates:
mean of x mean of y

11.08052 10.87926
X1 und X2 besitzen unterschiedliche Mittelwert.

c) Wilcoxon rank sum test with continuity correction
data: X1 and X2
W = 4165, p-value = 0.04145
alternative hypothesis: true location shift is not equal to 0
X1 und X2 besitzen unterschiedliche Median, weil p<0.05.

5)a) Zuerst muss man die Datei PISA.csv von Ilias hochladen. Danach muss man die Datei mit read.table() Befehl lesen. Man nutzt auch attach-Befehl um einfacher mit den Daten umgehen zu können.

b) Man kann schon merken, dass sich der PISA-Score zur Lese und Kompetenz in der Mathematik verringert hat. Im Gegensatz dazu hat sich der mittlere PISA-Score in den Naturwissenschaften erhöht.



c) Man macht t.test(X,Y,paired=TRUE,alternative=”greater”/”less”) in R Befehl.
Paired t-test
data: R00 and R06

t = 2.2964, df = 51, p-value = 0.0129
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
1.539612 Inf
sample estimates:
mean difference
5.692308

Paired t-test
data: M00 and M06
t = -0.008081, df = 51, p-value = 0.5032
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
-4.00601 Inf
sample estimates:
mean difference
-0.01923077

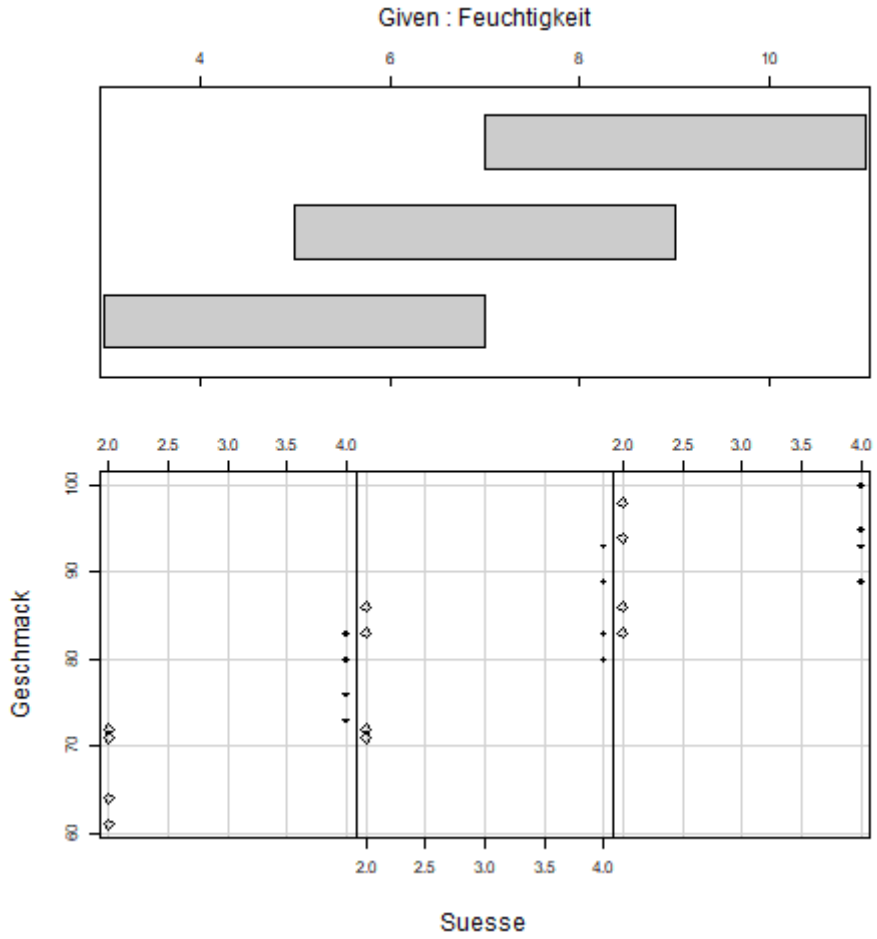
Paired t-test
data: S00 and S06
t = -1.2842, df = 51, p-value = 0.1024
alternative hypothesis: true mean difference is less than 0
95 percent confidence interval:
-Inf 0.8844165
sample estimates:
mean difference
-2.903846

Die Lesekompetenz hat sich signifikant verschlechtert, da p-Wert 0.0129 < 0.05 ist. Die Mathekompetenz und Naturwissenschaftenkompetenz nicht signifikant verändert, weil p-Wert 0.5032 > 0.05 und p-Wert 0.1024 > 0.05 sind.

6) Zuerst muss man die Datei Hustensaft.csv von Ilias hochladen. Danach muss man die Datei mit read.table() Befehl lesen. Man nutzt auch attach-Befehl um einfacher mit den Daten umgehen zu können. Danach nutzt man t.test(x,mu,alternative) Befehl in R.
One Sample t-test
data: Kon
t = -1.7586, df = 8, p-value = 0.05835
alternative hypothesis: true mean is less than 40
95 percent confidence interval:
-Inf 40.04593
sample estimates:
mean of x
39.2
p-Wert = 0.05835, H0 kann nicht abgelehnt werden. Die Durchschnittskonzentration von Halsruhe in den entnommen Flaschen ist nicht signifikant von der Sollkonzentration von 40 g/Liter abweichend und begründet keinen Produktionsstopp.

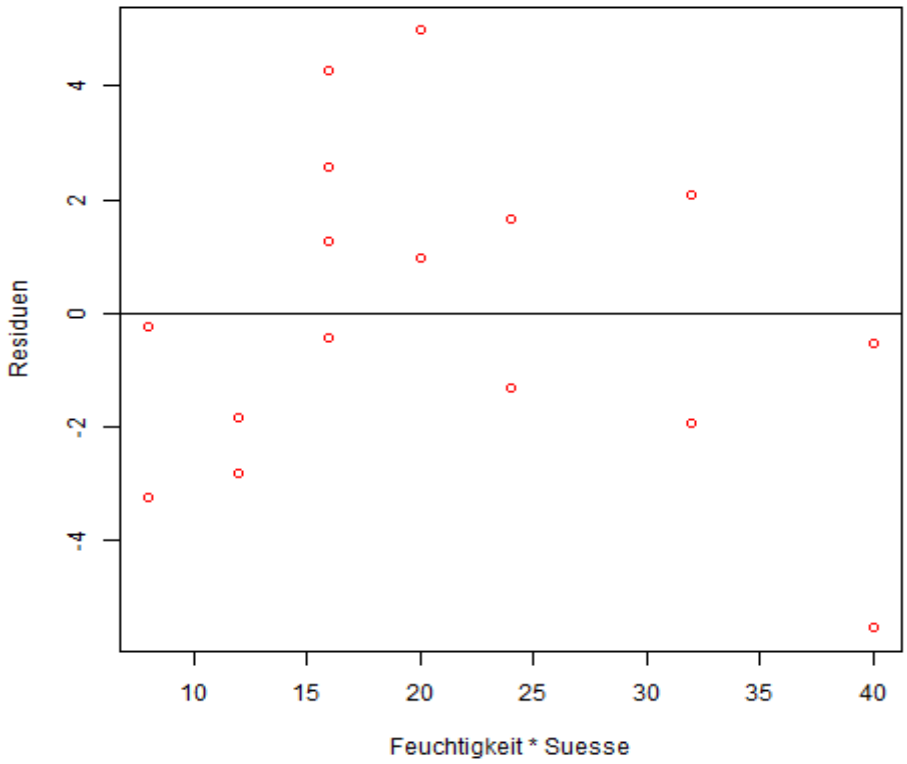
7)a) Zuerst muss man die Datei Suess.csv von Ilias hochladen. Danach muss man die Datei mit read.table() Befehl lesen. Man nutzt auch attach-Befehl um einfacher mit den Daten umgehen zu können. Danach abspeichert man die Datei in einem Dataframe mit data.frame() Befehl in R.

b) In alle gegebene Feuchtigkeit sind, wenn mehr süße gibt, dann das Geschmack ist besser.



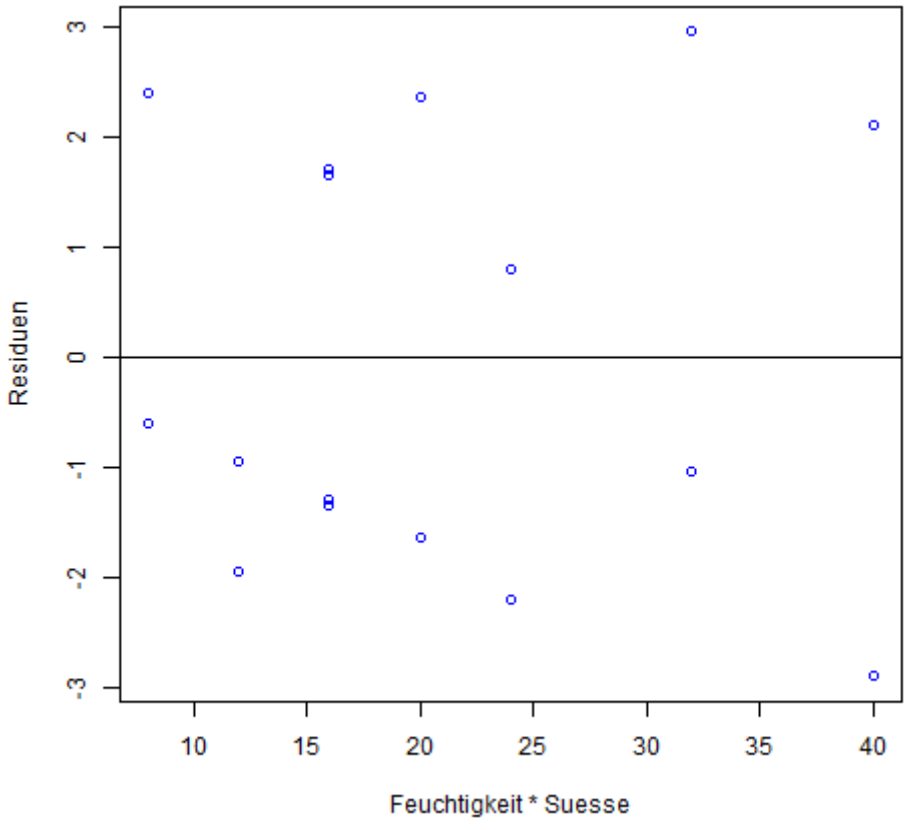
c)Man fittet das Modell mit lm() Befehl in R.

d)Es ist Varianzhomogenität und es ist linear.

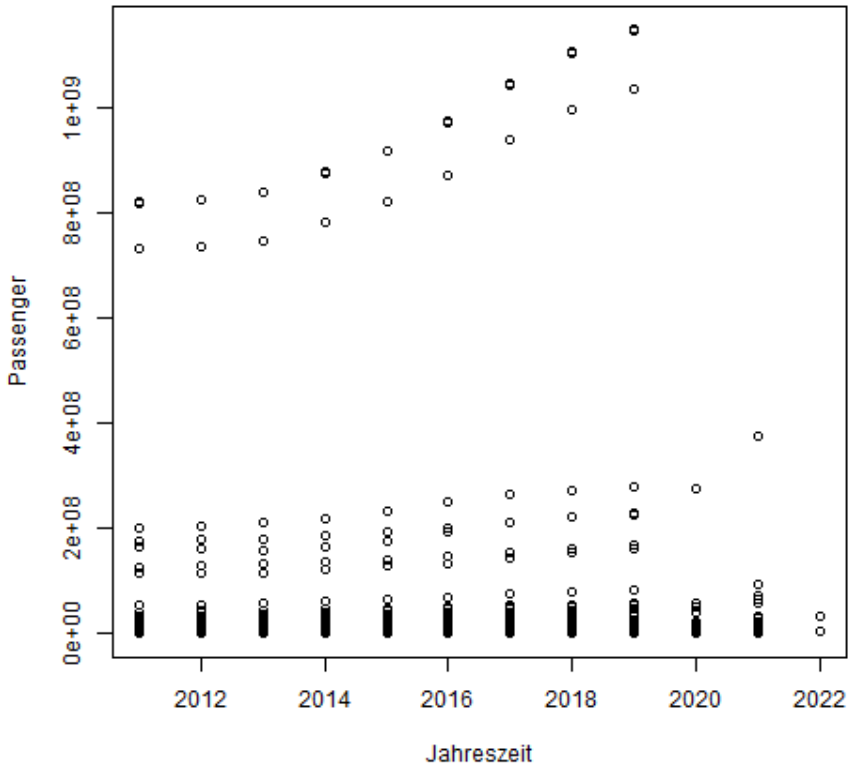


e)Man macht model2 mit lm() Befehl in R.

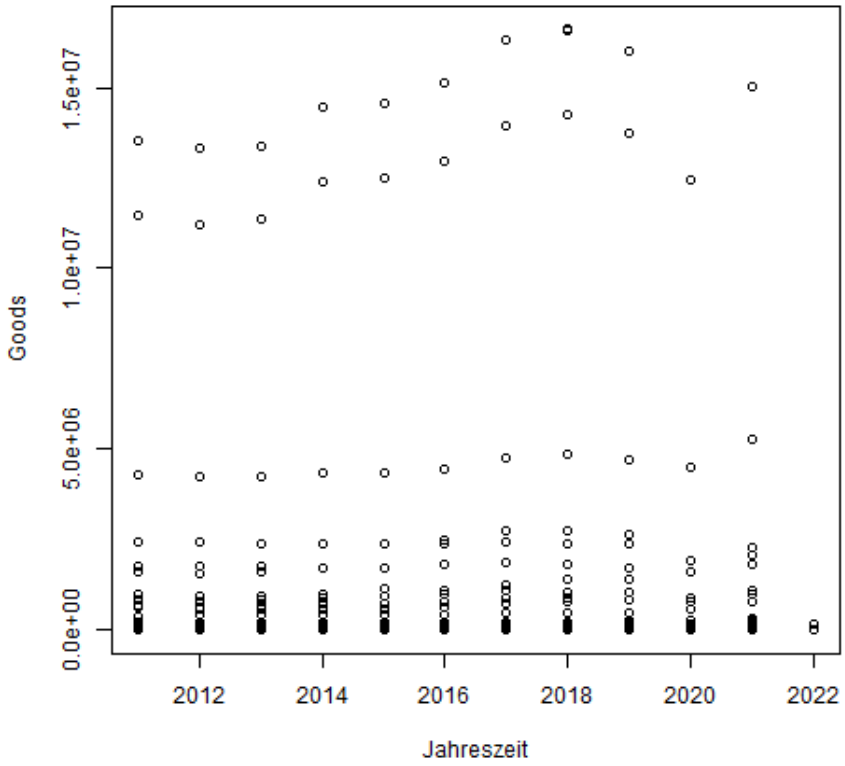
f)Die p-Werte ist 8.469e-10 < 0.05 heißt, dass die Anpassung des Modells an die Daten sich verbessert und der Koeffizient β3 der Interaktion signifikant von 0 verschieden ist. Die Residuen Werte in aufgabe f mehr sicher als in aufgabe d.



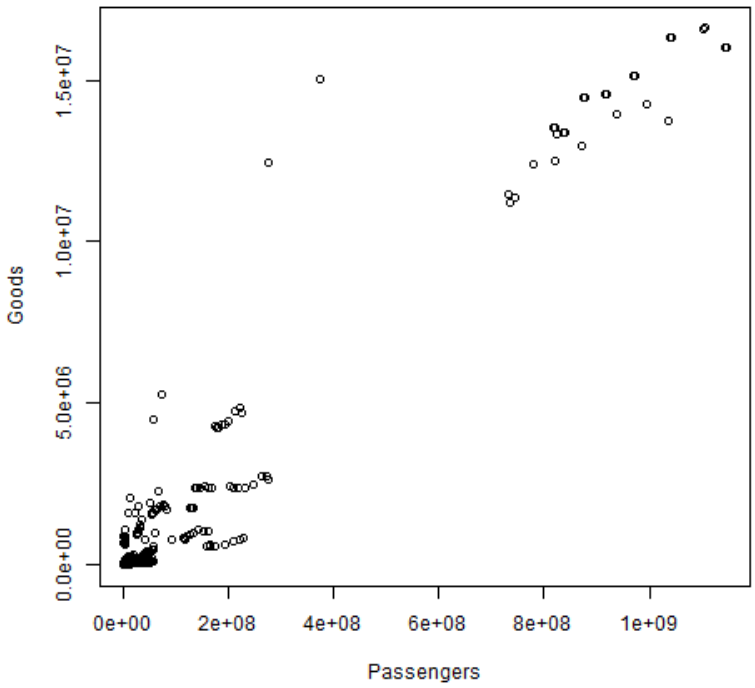
8)a) Zuerst muss man die Datei von Webseite hochladen. Danach muss man die Datei mit read.table() Befehl lesen. Danach nutzt man plot() Befehl für die Visualisierung in R. Je größer die Jahreszeit, desto höher die Anzahl der Passenger. Passenger und Jahreszeit abhängig miteinander.



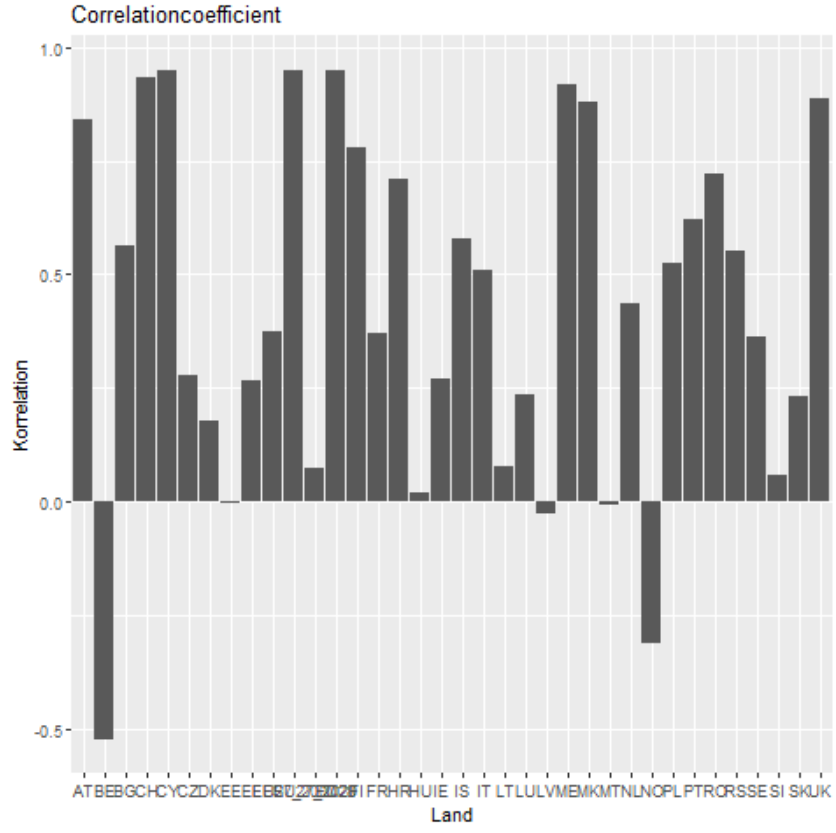
Von 2011 bis 2018 die Anzahl der Goods immer vergrößert wird und danach seit 2018 bis 2022 die Anzahl der Goods immer verringert. 2018 ist die maximale Anzahl der Goods. Goods und Jahreszeit abhängig miteinander.



b)Je größer die Anzahl der Passengers, desto größer die Anzahl der Goods. A und B abhängig miteinander. A und B korreliert miteinander, weil die Korrelationskoeffizient 0. 9686908 ist. Die Korrelation ist signifikant, weil die p-Werte 2.866642e-236 < 0.05 ist.

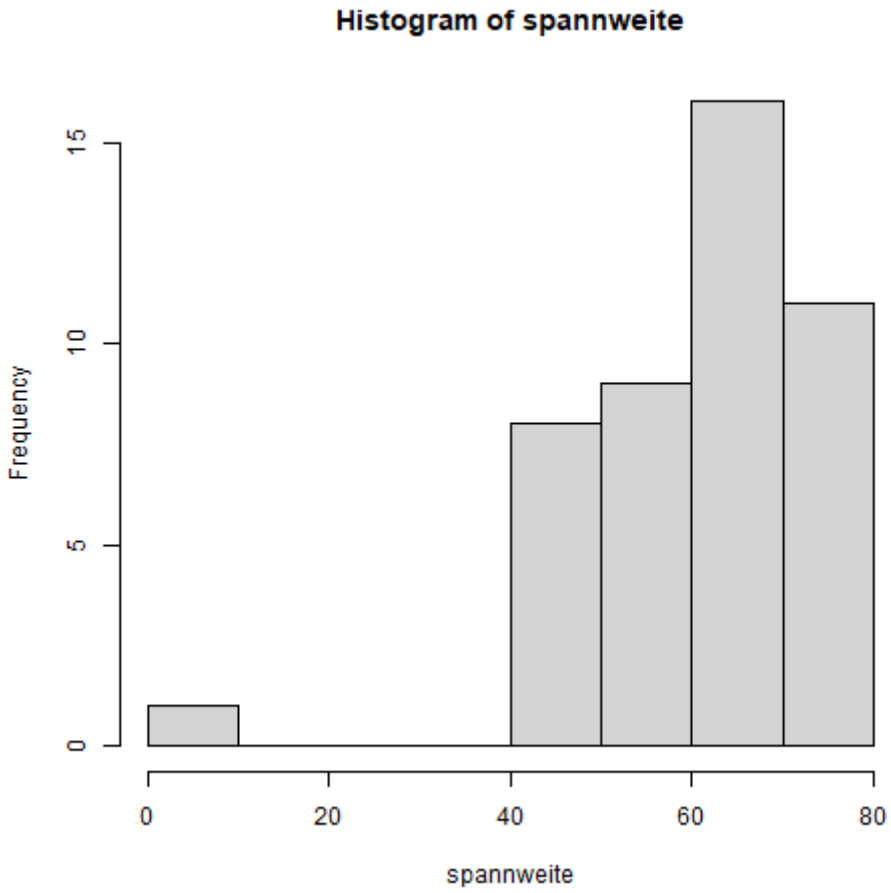


c)Land mit größte Korrelation ist EU27_2007(die 11.Land). Land mit kleinste Korrelation ist Belgien(die 2.Land). Man kann sehen in der ggplot, dass EU27_2007 stark positiv korreliert und Belgien negativ korreliert.

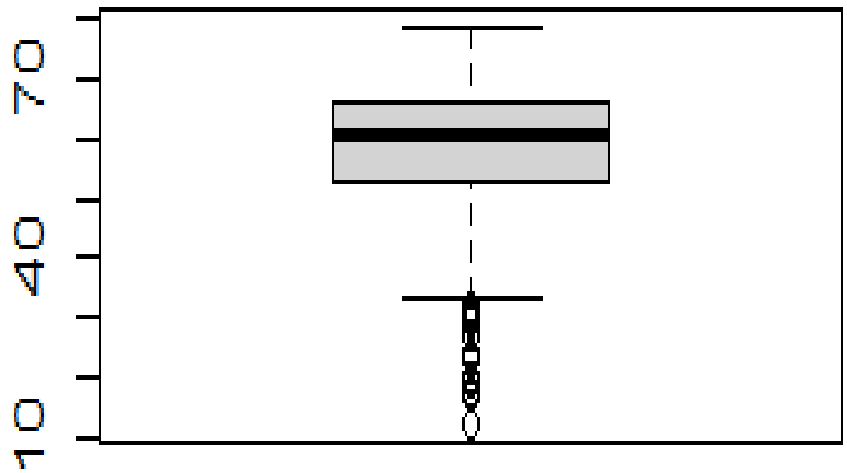


9)a)Zuerst liest man den SEPCTF Datensatz mit Hilfe des load()-Befehls in R. Danach überprüft man die Information der R-Data mit head() Befehl in R und man checkt die Dimension der Daten mit dim() Befehl in R.

b)Zuerst macht man die Spannweite. Danach macht man die summary und die Histogramm. Von der Histogramm sieht man, dass die spannweite zwischen 60 bis 70 die höchste Frequenz hat. Es gibt nur eine Frequenz, dass die spannweite zwischen 1 bis 10 liegt.



c)Man sucht für die Varianz von der X0 bis X44. Man findet die höchste 6 Varianz von X0 bis X44. Die höchste 6 Varianz sind X42,X44,X41,X26,X30,X25. Es gibt viele Werte, die kleiner als die Lower Fence ist.



10)a) Kreuzvalidierungsverfahren sind auf Resampling basierende Testverfahren der Statistik, die z. B. im Data-Mining die zuverlässige Bewertung von Maschinen gelernten Algorithmen erlauben. Es wird unterschieden zwischen der einfachen Kreuzvalidierung, der stratifizierten Kreuzvalidierung und der Leave-One-Out-Kreuzvalidierung.

b)Man macht eine 10 fache Kreuzvalidierung mittels einer for-Schleife. Innerhalb for-Schleife macht man train.data, test.data, logistic_model, predicted, predicted_values, accuracy[i].

c) ROC CURVE
Sensitivity (True Positive Rate): 0.50000
Specificity (True Negative Rate): 0.92000
False Positive Rate: 1- Specificity = 0.08000
False Negative Rate: 1- Sensitivity = 0.50000
Accuracy : 0.8889

Konfidenzintervall
95% CI : (0.7084, 0.9765)

AUC Wert = 0.94

d)Um die AUC-Wert zu sichtbar, nutzt man print.auc=TRUE. Um die Konfidenzintervall zu sichtbar, nutzt man print.ci=TRUE. AUC ist die Fläche unter der Linie.