Serie 10

Aufgabe 10.1

Wir kommen nochmals auf die Datei `Diet.csv` zurück, in der 76 Personen aufgelistet, die jeweils einer der Diäten 1,2 oder 3 für 6 Wochen machten.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt


df = pd.read_csv("../../../Themen/Varianzanalyse/Uebungen_de/Daten/Diet.csv")

df["weight_loss"] = df["weight6weeks"] - df["pre.weight"]

df.head()
```

```
##    Person  gender  Age  Height  pre.weight  Diet  weight6weeks  weight_loss
## 0      25       0   41     171          60     2          60.0          0.0
## 1      26       0   32     174         103     2         103.0          0.0
## 2       1       0   22     159          58     1          54.2         -3.8
## 3       2       0   46     192          60     1          54.0         -6.0
## 4       3       0   55     170          64     1          63.3         -0.7
```

Es gibt neben der Diät noch einen weiteren Faktor nämlich **`gender`** mit Mann (0) und Frau (1). Wir wollen untersuchen, ob Männer und Frauen unterschiedlich auf die Diäten reagieren.

a) Erstellen Sie zunächst einen Boxplot und Stripchart (siehe letzte Übungsserie) mit **`weight.loss`** und **`gender`**. Wie interpretieren Sie diese?

b) Führen Sie folgenden Interaction-Plot aus und interpretieren Sie diesen.

```
from statsmodels.graphics.factorplots import interaction_plot

interaction_plot(x=df["gender"], trace=df["Diet"], response=df["weight_loss"])
```

c) Was geschieht, wenn Sie **`gender`** und **`Diet`** austauschen? Interpretieren Sie diesen Plot.

d) Führen Sie einen Anova-Hypothesentest ohne Wechselwirkung auf Signifikanzniveau von 5 % durch. Stellen Sie die beiden Nullhypothesen auf und interpretieren Sie die p-Werte.

e) Führen Sie den Test vorher noch mit Wechselwirkung durch. Interpretieren Sie wieder die entsprechenden Resultate.

Aufgabe 10.2

Die Daten `mathGender.dat` stammen aus einer Beobachtungsstudie um die Beziehung zwischen dem Resultat beim ACT Math Usage Test und den beiden Variablen Geschlecht (1=female, 2=male) und Level der erbrachten Mathematikkurse (1=algebra only, 2=algebra+geometry, 3=through calculus) zu untersuchen.

Es wurden 861 high school seniors untersucht. Die Resultate, ACT score, gehen von 0 to 36 mit einem Median von 15 und einem Durchschnitt von 15.33.

Untersuchen Sie wie in Aufgabe 1 die Beziehung von ACT score und dem Geschlecht und den mathematischen Vorkenntnissen. Verwenden Sie dazu Boxplots, Stripcharts, Interaction-Plots und Hypothesentests. Interpretieren Sie jeweils Ihre Resultate.

Lesen Sie Datei folgendermassen ein:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv(r"*mathGender.dat", sep=" ")

df.head()
```

Der $*$ steht wie immer für ihren Pfad.

Aufgabe 10.3

Eine Konsumentenschutzorganisation lässt den jährlichen Energieverbrauch von fünf verschiedenen Marken von Luftentfeuchtern vergleichen. Weil der Energieverbrauch von der aktuellen Luftfeuchtigkeit abhängt, wurde jede Marke bei vier verschiedenen Luftfeuchtigkeitsniveaus im Bereich moderat bis hohe Feuchtigkeit getestet. Von jeder Marke wurden deshalb vier Geräte zufällig den Luftfeuchtigkeitsniveaus zugeordnet. Der aus dem Versuch resultierende Energieverbrauch (in kWh) ist in folgender Tabelle festgehalten:

a) Tippen Sie die Daten selber ein und lesen Sie sie in **Python** ein.
Python-Hinweise: Die Daten werden in drei Spalten eingelesen: Spalte mit Energieverbrauch, mit Marke, mit Luftfeuchtigkeitsniveau.

```
from pandas import DataFrame
import pandas as pd
import numpy as np
import seaborn as sns
import scipy.stats as st
```

Marke	Luftfeuchtigkeitsniveau			
	1	2	3	4
1	685	792	838	875
2	722	806	893	953
3	733	802	880	941
4	811	888	952	1005
5	828	920	978	1023

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from statsmodels.graphics.factorplots import interaction_plot
import matplotlib.pyplot as plt
from patsy.contrasts import Sum

df=DataFrame({
 "Luftfeuchtigkeitsniveau" : np.tile(["1", "2", "3", "4"], 5),
  "Marke": np.repeat(["1", "2", "3", "4", "5"], 4),
  "Energieverbrauch" : np.array([685, 792, 838 , ..])
})
```

b) Können Sie auf dem 5 %-Signifikanzniveau schliessen, dass es zwischen den fünf verschiedenen Marken einen Unterschied im Energieverbrauch gibt?
Python-Hinweise:

```
fit = ols("Energieverbrauch ~ C(Marke, Sum) +
          C(Luftfeuchtigkeitsniveau, Sum)", data=df).fit()
anova_lm(fit)
```

c) Können Sie auf dem 1 % Signifikanzniveau schliessen, dass es zwischen den Niveaus (oder Stufen) des Block-Faktor `Luftfeuchtigkeit` einen Unterschied im Energieverbrauch gibt?

Stützt dieses Resultat den Entscheid der Versuchsleiterin, `Luftfeuchtigkeit` als Block-Faktor einzusetzen?

d) Überprüfen Sie mit einem Interaktionsplot die Additivität.

Könnte man mit diesen Daten auf Wechselwirkung testen?

Aufgabe 10.4

In drei Städten der USA (Variable `STADT`) wurde der Benzinverbrauch von 5 Automobil-Typen (Variable `AUTO`, Werte von 1 bis 5) ermittelt. Pro Kombination wurden 3 Testfahrten durchgeführt. Die Zielgrösse (Variable `KMP4L`) ist die Strecke in km, welche mit 4 Litern Benzin zurückgelegt werden konnte. Die Daten befinden sich im Datensatz `automob.dat`

a) Stellen Sie die Variablen `AUTO` und `KMP4L` graphisch dar (wenn möglich mit unterschiedlichen Symbolen oder Farben bezüglich der Variablen `STADT`).

   Python-Hinweise:

```
from pandas import DataFrame
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as st

automob = pd.read_csv(".../automob.dat", sep=" ")
df = DataFrame(automob)
sns.stripplot(x="STADT", y="KMP4L", hue="AUTO", jitter=True, data=automob)
```

b) Analysieren Sie die Zielgrösse `KMP4L` mit einer zweifaktoriellen Varianzanalyse. Verwenden Sie das volle Modell mit Interaktion. **Python**-Hinweise:

```
from pandas import DataFrame
import pandas as pd
import numpy as np
import seaborn as sns
import scipy.stats as st
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from statsmodels.graphics.factorplots import interaction_plot
import matplotlib.pyplot as plt

fit = ols("KMP4L ~ C(STADT,Sum)*C(AUTO,Sum)", data=automob).fit()
anova_lm(fit)
```

c) Stellen Sie die Zellenmittelwerte als Interaktionsplot dar. Wie lässt sich die signifikante Wechselwirkung erklären?
   Python-Hinweis:

```
automob = pd.read_table(".../automob.dat", sep=" ")
df = DataFrame(automob)

df.reset_index(inplace=True)
interaction_plot(x=df["..."], trace=df["..."], response=df["..."])
```

d) Führen Sie für jede Stadt separat eine einfaktorielle Varianzanalyse durch.

e) Wiederholen Sie Teilaufgabe b) ohne die Werte von San Francisco.

Aufgabe 10.5

a) Der Datensatz `stream` enthält die Zinkstufen (Variable **`ZINC`**) verschiedener Flüsse (Variable **`STREAM`**) und die entsprechende Biodiversität (Variable **`DIVERSITY`**). Zusätzlich kodiert die Variable **`ZNGROUP`** die verschiedenen Zink-Gruppen numerisch. Wir wollen untersuchen, ob eine signifikante Beziehung zwischen der Biodiversität und den Zink-Gruppen besteht.

   Lesen Sie den in der Datei `stream.dat` gespeicherten Datensatz ein. Erstellen Sie einen Boxplot und Stripchart von **`DIVERSITY`** versus **`ZNGROUP`**, und kommentieren Sie die Graphik in Bezug auf Unterschiede und Ausreisser.

b) Sie möchten feststellen, ob es einen signifikanten Unterschied der **`DIVERSITY`** für die unterschiedlichen Zink-Gruppen gibt. Formulieren Sie ein entsprechendes Modell, und führen Sie die entsprechende Varianzanalyse durch.

c) Wie lauten die Schätzungen der Gruppenmittelwerte? Sind diese kompatibel mit Ihrer Beobachtung der Stripcharts?

d) Nun möchten Sie überprüfen, ob die unterschiedlichen Flüsse **`STREAM`** neben der Zinkgruppe **`ZNGROUP`** einen Einfluss auf **`DIVERSITY`** haben. Führen Sie eine Zweiweg-Varianzanalyse durch. Wie interpretieren Sie die Variable **`STREAM`** in Bezug auf den Versuchsplan?

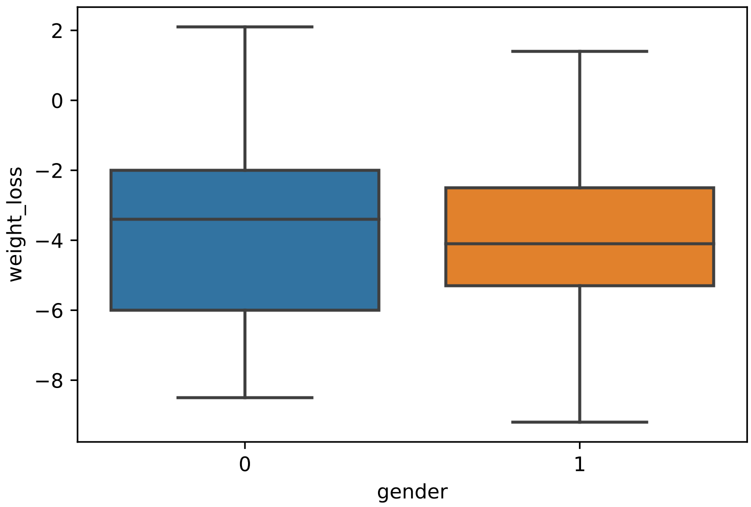
Kurzlösungen vereinzelter Aufgaben

Musterlösungen zu Serie 10

Lösung 10.1

a) Boxplot:

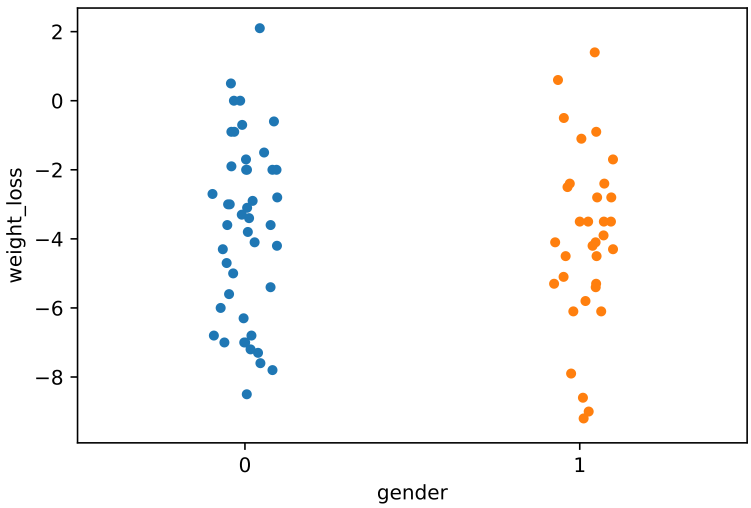
```
sns.boxplot(x="gender", y="weight_loss", data=df)
```



Kaum ein Unterschied zwischen den Geschlechtern erkennbar. Beide haben in etwa denselben Median. Das Geschlecht hat also scheinbar auf den Gewichtsverlust keinen Einfluss.

Stripchart:

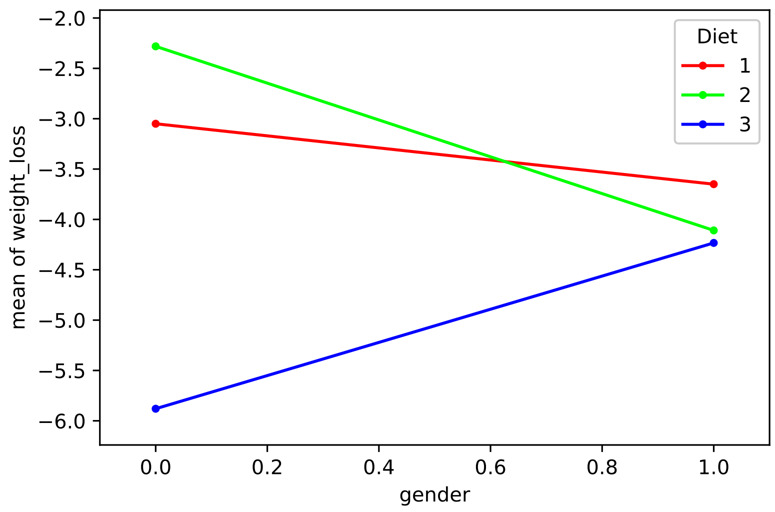
```
sns.stripplot(x="gender", y="weight_loss", data=df)
```



Auch hier kaum ein Unterschied erkennbar.

b) Interaction-Plot:

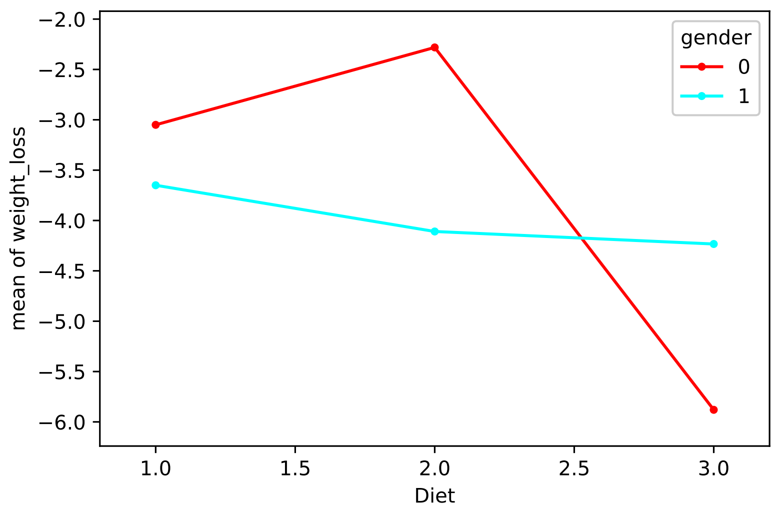
```
from statsmodels.graphics.factorplots import interaction_plot

interaction_plot(x=df["gender"], trace=df["Diet"], response=df["weight_loss"])
```



Bei Diät 1 und 2 nehmen die Frauen stärker ab als die Männer. Bei Diät 3 ist es gerade umgekehrt. Es *scheint* hier eine Wechselwirkung zu geben.

c) Plot: Interaction-Plot:

```
interaction_plot(x=df["Diet"], trace=df["gender"], response=df["weight_loss"])
```



Bei den Frauen führen alle Diäten zu einem ähnlichen durchschnittlichen Gewichtsverlust. Bei den Männern ist aber Diät 3 wesentlich effektiver als die beiden anderen Diäten.

d) Es gibt zwei Nullhypothesen:

- Männer und Frauen denselben durchschnittlichen Gewichtsverlust
- Alle drei Diäten führen zum gleichen durchschnittlichen Gewichtsverlust

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from patsy.contrasts import Sum

fit = ols("weight_loss~gender+Diet", data=df).fit()
anova_lm(fit)
```

```
##              df      sum_sq    mean_sq         F    PR(>F)
## gender      1.0    1.658524   1.658524  0.273855  0.602300
## Diet        1.0   45.397745  45.397745  7.496050  0.007719
## Residual   75.0  454.216680   6.056222       NaN       NaN
```

Der p-Wert für **`gender`** ist 0.602 und damit wesentlich grösser als das Signifikanzniveau. Die Nullhypothese wird somit *nicht* verworfen. Der durchschnittliche Gewichtsverlust ist bei Frauen und Männer gleich. Dies ist nach dem Boxplot in a) auch nicht weiter überraschend.

Der p-Wert für **`Diet`** ist 0.0077 und damit deutlich unter dem Signifikanzniveau. Die Nullhypothese wird somit verworfen. Die Diäten sind statistisch signifikant nicht alle gleich wirksam.

e) Es gibt drei Nullhypothesen:

- Männer und Frauen denselben durchschnittlichen Gewichtsverlust
- Alle drei Diäten führen zum gleichen durchschnittlichen Gewichtsverlust
- Geschlecht und Diäten zeigen keine Wechselwirkung. Das heisst, Männer und Frauen reagieren gleich auf die entsprechenden Diäten.

```
fit = ols("weight_loss~gender*Diet", data=df).fit()
anova_lm(fit)
```

```
##                 df      sum_sq    mean_sq         F    PR(>F)
## gender         1.0    1.658524   1.658524  0.280477  0.597974
## Diet           1.0   45.397745  45.397745  7.677308  0.007067
## gender:Diet    1.0   16.637105  16.637105  2.813536  0.097692
## Residual      74.0  437.579575   5.913237       NaN       NaN
```

Der p-Wert für `gender` ist 0.598 und damit wesentlich grösser als das Signifikanzniveau. Die Nullhypothese wird somit *nicht* verworfen. Der durchschnittliche Gewichtsverlust ist bei Frauen und Männer gleich. Dies ist nach dem Boxplot in a) auch nicht weiter überraschend.

Der p-Wert für `Diet` ist 0.007 und damit deutlich unter dem Signifikanzniveau. Die Nullhypothese wird somit verworfen. Die Diäten sind statistisch signifikant nicht alle gleich wirksam.

Der p-Wert für `gender:Diet` ist 0.097 und liegt über dem Signifikanzniveau. Die Nullhypothese wird somit *nicht* verworfen. Es gibt keine statistisch signifikante Wechselwirkung. Männer und Frauen reagieren also gleich auf die jeweiligen Diäten.

Der Interaction-Plot in b) suggeriert zwar eine Wechselwirkung, aber sie ist nicht signifikant.

Lösung 10.2

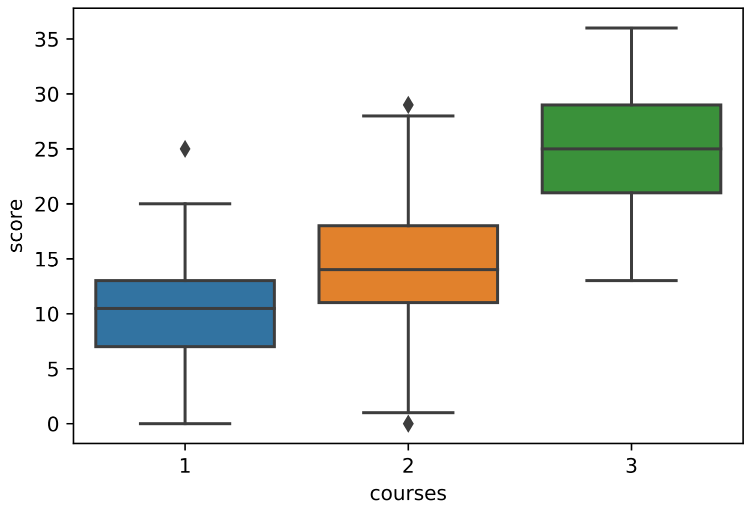
Einlesen der Datei:

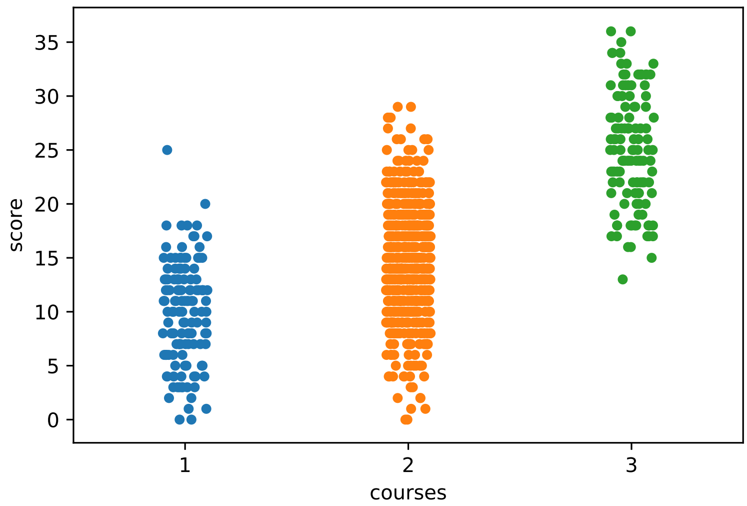
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv("../../../Themen/Varianzanalyse/Uebungen_de/Daten/mathGender.dat", sep=" ")

df.head()


##    score  courses  gender
## 0      5        1       2
## 1     13        1       2
## 2      7        1       2
## 3     20        1       2
## 4     11        1       2
```

```
df <- read.table("../../../Themen/Varianzanalyse/Uebungen_de/Daten/mathGender.da
    header = T, sep = "")
head(df)


##   score courses gender
## 1     5       1      2
## 2    13       1      2
## 3     7       1      2
## 4    20       1      2
## 5    11       1      2
## 6    16       1      2
```

Boxplot und Stripchart für **`courses`**

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(x="courses", y="score", data=df)
sns.stripplot(x="courses", y="score", data=df)
```

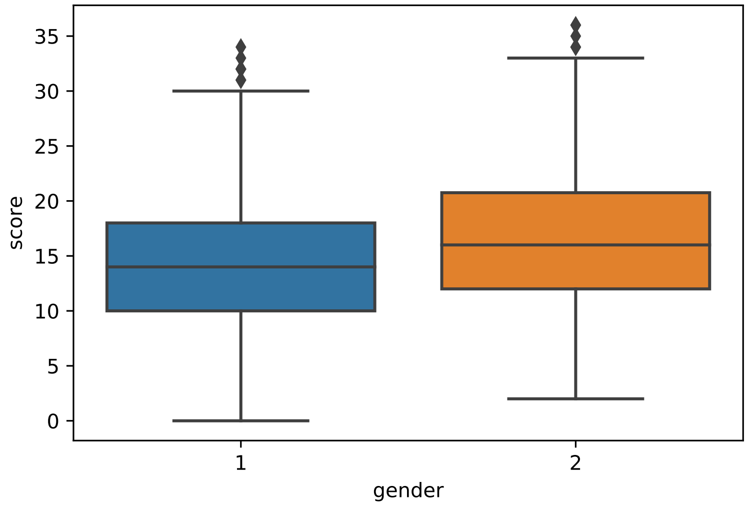


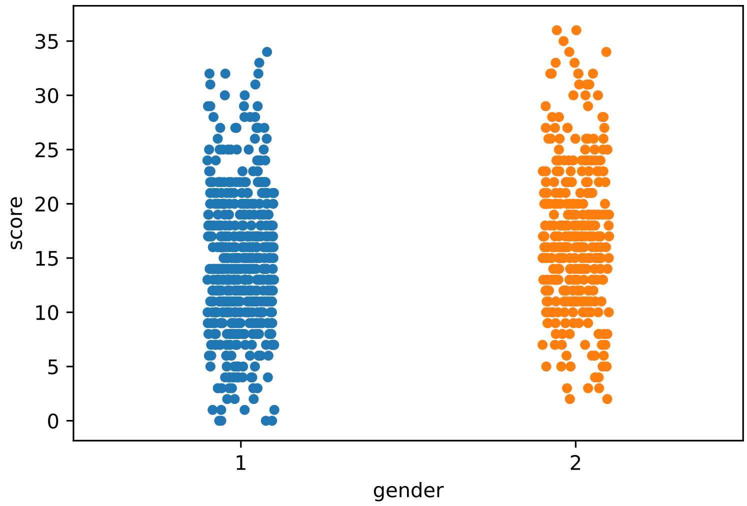


Es scheint in der Tat Unterschiede in den Resultaten zu geben. Dies ist aber nicht weiter erstaunlich, da die Studierenden mit mehr mathematischen Vorkenntnissen bevorteilt sind. Sind die Unterschiede aber signifikant?

Boxplot und Stripchart für **`gender`**

```
sns.boxplot(x="gender", y="score", data=df)
sns.stripplot(x="gender", y="score", data=df)
```

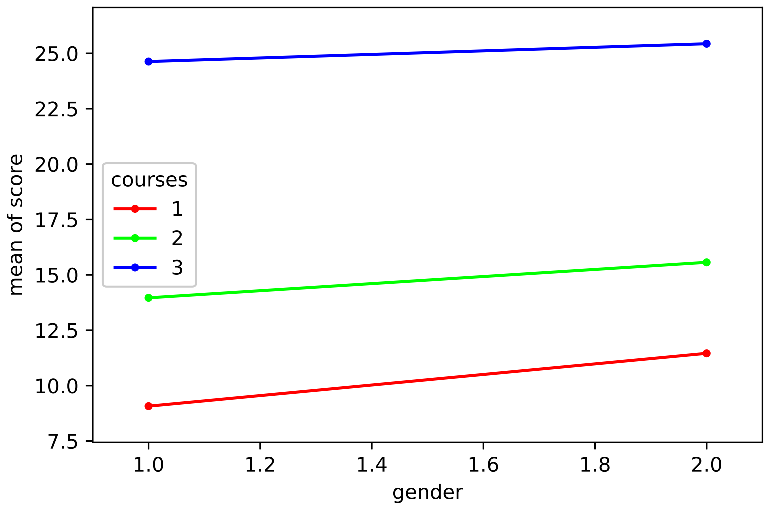




Hier scheinen die Unterschiede weniger markant. Die Frauen haben einen etwas schlechteren Median und Durchschnitt. Aber auch hier: Sind die Unterschiede signifikant?

Interaktion-Plot:

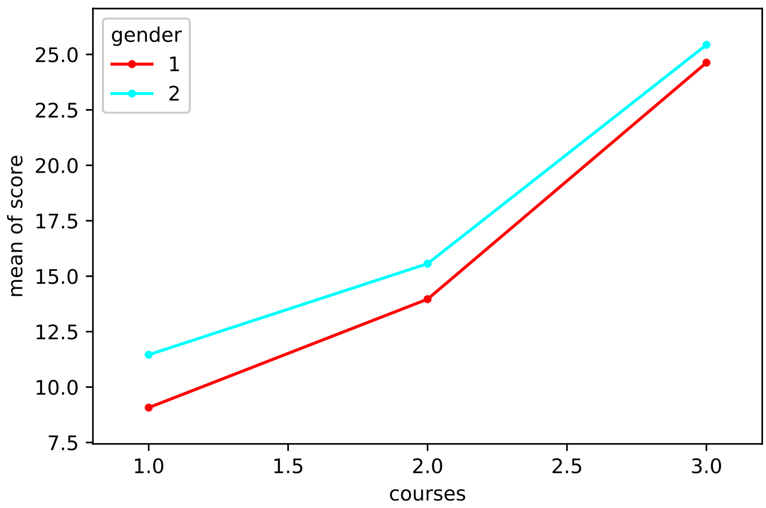
```
from statsmodels.graphics.factorplots import interaction_plot

interaction_plot(x=df["gender"], trace=df["courses"], response=df["score"])
```



Unabhängig von den Vorkenntnissen schneiden die Männer leicht besser ab, als die Frauen. Es gibt aber offensichtlich Unterschiede, wenn man die Vorkenntnisse betrachtet werden. Die Linien liegen weit auseinander.

Die Linien sind hier parallel und somit scheint keine Wechselwirkung vorhanden zu sein.

```
from statsmodels.graphics.factorplots import interaction_plot

interaction_plot(x=df["courses"], trace=df["gender"], response=df["score"])
```



Auch hier sehen wir, dass die Gruppenmittelwerte der drei Levels sehr unterschiedlich sind. Desweiteren sehen wir, dass je grösser die Vorkenntnisse sind, die Resultate der

Frauen und Männer annähern. Ob diese Wechselwirkung signifikant ist, lässt sich hier nicht erkennen.

Hypothesentest ohne Wechselwirkung:

Nullhypothesen:

- Alle 3 Levels an Vorkenntnissen schneiden gleich gut ab
- Frauen und Männer schneiden gleich gut ab

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from patsy.contrasts import Sum

fit = ols("score~courses+gender", data=df).fit()
anova_lm(fit)
```

```
##                df         sum_sq        mean_sq           F         PR(>F)
## courses       1.0   14056.755673   14056.755673  540.656992   3.950023e-93
## gender        1.0     685.427351     685.427351   26.363202   3.501577e-07
## Residual    858.0   22307.482481      25.999397         NaN            NaN
```

Der p-Wert für `courses` ist $3.95 \cdot 10^{-93}$ und damit weit unter dem Signifikanzniveau. Die Nullhypothese wird somit verworfen. Die Vorkenntnisse spielen eine statistisch signifikante Rolle zur Erreichung einer hohen Punktezahl. Dies ist nach dem Boxplot weiter oben auch nicht weiter überraschend.

Der p-Wert für `gender` ist $3.5 \cdot 10^{-7}$ und damit weit unter dem Signifikanzniveau. Die Nullhypothese wird somit verworfen. Die Resultate der Geschlechter sind statistisch signifikant unterschiedlich. Das kommt ein bisschen überraschend, da die Boxplots weiter oben relative nahe zusammen sind. Der Grund dafür liegt in der hohen Anzahl von Teilnehmenden, womit die schon ein relativ kleiner Unterschied signifikant wird.

Hypothesentest mit Wechselwirkung:

Nullhypothesen:

- Alle 3 Levels an Vorkenntnissen schneiden gleich gut ab
- Frauen und Männer schneiden gleich gut ab
- Es gibt keine Wechselwirkung zwischen Geschlecht und dem Level der Vorkenntnisse. Die Unterschiede der Resultate innerhalb der Levels sind jeweils gleich.

```
fit = ols("score~courses*gender", data=df).fit()
anova_lm(fit)
```

```
##                   df        sum_sq       mean_sq           F         PR(>F)
## courses          1.0  14056.755673  14056.755673  540.028131  5.044214e-93
## gender           1.0    685.427351    685.427351   26.332538  3.556793e-07
## courses:gender   1.0      0.052701      0.052701    0.002025  9.641209e-01
## Residual       857.0  22307.429780     26.029673         NaN           NaN
```

Der p-Wert für **courses** wie oben.

Der p-Wert für **gender** wie oben.

Der p-Wert für **courses:gender** ist 0.96 und somit weit über Signifikanzniveau. Die Nullhypothese wird also nicht verworfen. Es gibt keine Wechselwirkung.

Lösung 10.3

a) (zu **R**)

```
from pandas import DataFrame
import pandas as pd
import numpy as np
import seaborn as sns
import scipy.stats as st
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from statsmodels.graphics.factorplots import interaction_plot
import matplotlib.pyplot as plt
from patsy.contrasts import Sum

df=DataFrame({
 "Luftfeuchtigkeitsniveau" : np.tile(["1", "2", "3", "4"], 5),
 "Marke": np.repeat(["1", "2", "3", "4", "5"], 4),
 "Energieverbrauch" : np.array([685, 792, 838, 875, 722, 806, 893,
953,
                        733, 802, 880, 941, 811, 888, 952, 1005,
                        828, 920, 978, 1023])
})
df.head()
```

```
##   Luftfeuchtigkeitsniveau Marke  Energieverbrauch
## 0                       1     1               685
## 1                       2     1               792
## 2                       3     1               838
## 3                       4     1               875
## 4                       1     2               722
```

b) (zu **R**)

```
fit = ols("Energieverbrauch ~ C(Marke, Sum) + C(Luftfeuchtigkeitsniveau, Sum)", data=df).fit()
anova_lm(fit)
```

```
##                                    df     sum_sq  ...          F        PR(>F)
## C(Marke, Sum)                     4.0   53231.00  ...   95.567325  5.419353e-09
## C(Luftfeuchtigkeitsniveau, Sum)   3.0  116217.75  ...  278.199282  2.363880e-11
## Residual                         12.0    1671.00  ...         NaN           NaN
##
## [3 rows x 5 columns]
```

Da der p-Wert von $5.419 \cdot 10^{-9}$ kleiner als das Niveau von 5 % ist, wird die Nullhypothese, dass alle α (Stufen zur Faktorvariable Marke) null sind, verworfen. Wir schliessen also daraus, dass es im Energieverbrauch Unterschiede gibt.

c) (zu **R**)

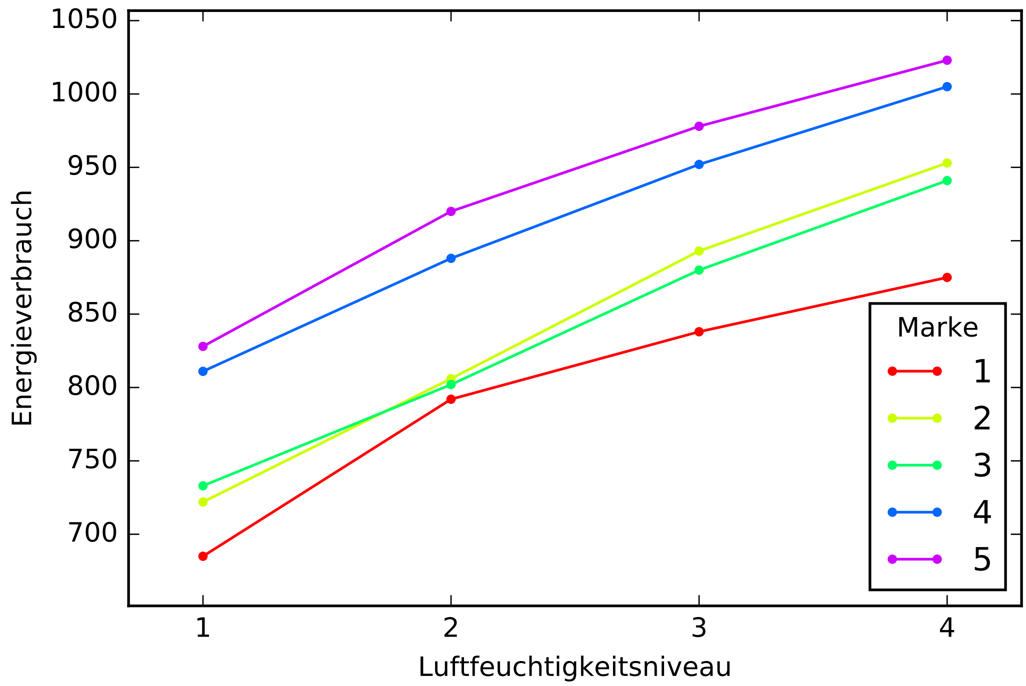
```
fit = ols("Energieverbrauch ~ C(Marke, Sum) + C(Luftfeuchtigkeitsniveau, Sum)", data=df).fit()
anova_lm(fit)
```

```
##                                    df     sum_sq  ...          F        PR(>F)
## C(Marke, Sum)                     4.0   53231.00  ...   95.567325  5.419353e-09
## C(Luftfeuchtigkeitsniveau, Sum)   3.0  116217.75  ...  278.199282  2.363880e-11
## Residual                         12.0    1671.00  ...         NaN           NaN
##
## [3 rows x 5 columns]
```

Da der P-Wert von $2.364 \cdot 10^{-11}$ kleiner als das Niveau von 1 % ist, wird auch die Nullhypothese, dass alle β (Stufen zur Faktorvariable Luftfeuchtigkeitsniveau) null sind, verworfen. Wir schliessen also daraus, dass es wichtig war, `Luftfeuchtigkeit` als Blockvariable einzusetzen

d) Wir erzeugen den Interaktionsplot wie folgt (zu **R**) :

```
interaction_plot(x = df["Luftfeuchtigkeitsniveau"], trace = df["Marke"],
    response = df["Energieverbrauch"])
plt.ylabel("Energieverbrauch")
plt.show()
```



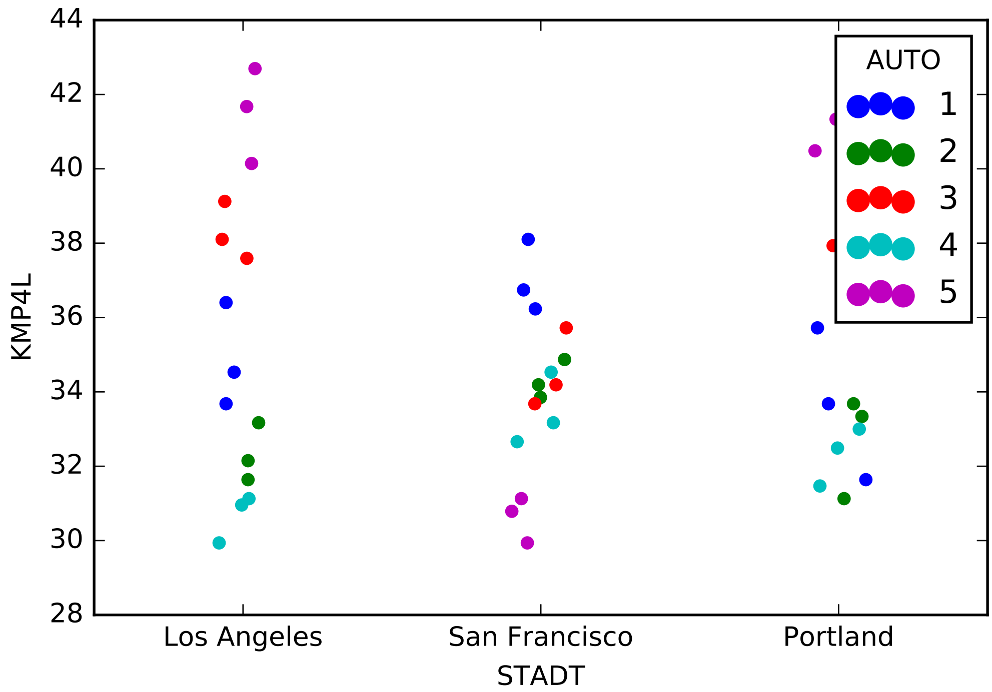
Da die Linien mehr oder weniger parallel verlaufen, kann man gut von der Additivität ausgehen. Es ist auch zu sehen, dass der Energieverbrauch mit steigender Luftfeuchtigkeit zunimmt.

Wechselwirkung kann nicht formal getestet werden, da wir keine Messwiederholungen haben, d.h. wir haben zu wenige Beobachtungen.

Lösung 10.4

a) (zu **R**)

```
automob = pd.read_csv("./Daten/automob.dat", sep = " ")
df = DataFrame(automob)
df.columns
sns.stripplot(x = "STADT", y = "KMP4L", hue = "AUTO", jitter = True,
    data = automob)
```



b) Die Anova-Tabelle erstellt sich mit **Python** wie folgt : (zu **R**)

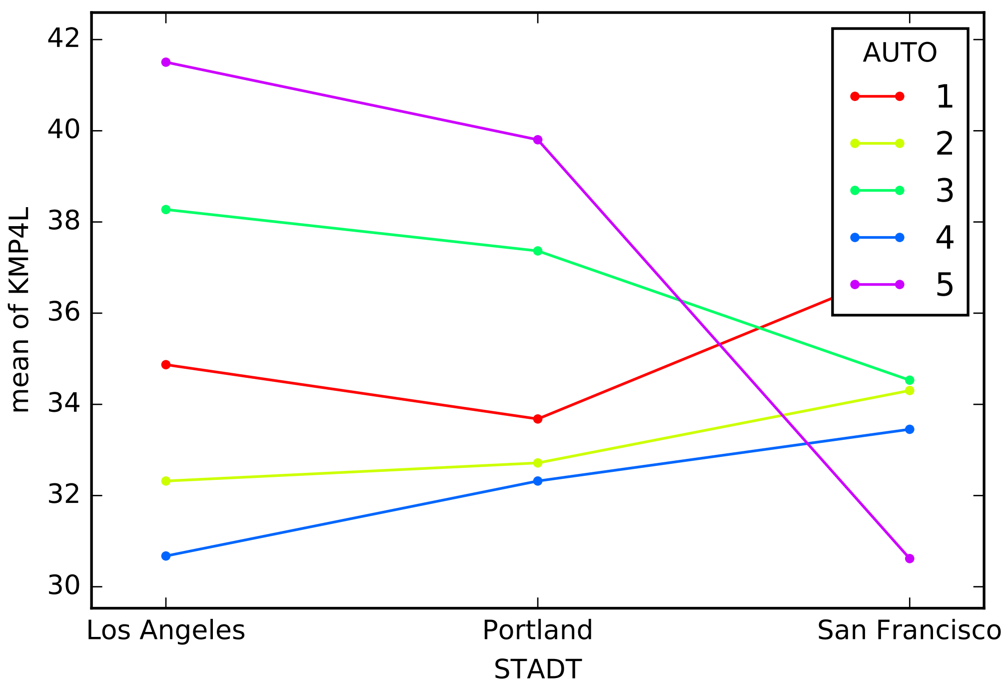
```
from pandas import DataFrame
import pandas as pd
import numpy as np
import seaborn as sns
import scipy.stats as st
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from statsmodels.graphics.factorplots import interaction_plot
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
automob = pd.read_csv("./Daten/automob.dat", sep=" ")
df = DataFrame(automob)
fit = ols("KMP4L ~ C(STADT,Sum)*C(AUTO,Sum)", data=automob).fit()
anova_lm(fit)

##                              df      sum_sq  ...          F        PR(>F)
## C(STADT, Sum)               2.0   19.599255  ...   7.438263  2.379745e-03
## C(AUTO, Sum)                4.0  179.741928  ...  34.107613  9.163330e-11
## C(STADT, Sum):C(AUTO, Sum)  8.0  244.619694  ...  23.209370  7.578447e-11
## Residual                   30.0   39.523858  ...        NaN           NaN
##
```

```
## [4 rows x 5 columns]
```

Sowohl die Faktoren STADT und AUTO wie auch die Wechselwirkung STADT:AUTO sind in diesem Modell auf dem 5 % Niveau signifikant.

c) Der Interaktions-Plot lässt sich mit **Python** wie folgt erstellen : (zu **R**)

```
automob = pd.read_table("./Daten/automob.dat", sep = " ")
df = DataFrame(automob)
df.reset_index(inplace = True)

interaction_plot(x = df["STADT"], trace = df["AUTO"], response = df["KMP4L"
```



Im Interaktion-Plot sind die Wechselwirkungen zwischen STADT und AUTO deutlich in Form von sich überschneidenden Linien zu sehen. Besonders auffällig ist das Verhalten in der Stadt San Francisco, welches ein anderes Verhalten zeigt als die beiden anderen Städten.

d) Die Anova-Tabellen für die einzelnen Städte lassen sich wie folgt anfertigen : (zu **R**)

```
fit_1 = ols("KMP4L ~ C(AUTO,Sum)", data=df[df["STADT"]=="Portland"]).fit()
anova_lm(fit_1)
```

```
##                  df       sum_sq    mean_sq          F    PR(>F)
## C(AUTO, Sum)    4.0   127.977055  31.994264  14.678319  0.000344
## Residual       10.0    21.796954   2.179695        NaN       NaN
```

```
fit_2 = ols("KMP4L ~ C(AUTO,Sum)", data=df[df["STADT"]=="San Francisco"]).fit()
anova_lm(fit_2)
```

```
##                  df      sum_sq    mean_sq          F    PR(>F)
## C(AUTO, Sum)    4.0   63.782132  15.945533  21.869048  0.000062
## Residual       10.0    7.291371   0.729137        NaN       NaN
```

```
fit_3 = ols("KMP4L ~ C(AUTO,Sum)", data=df[df["STADT"]=="Los Angeles"]).fit()
anova_lm(fit_3)
```

```
##                  df       sum_sq    mean_sq         F        PR(>F)
## C(AUTO, Sum)    4.0   232.602435  58.150609  55.72366  8.443631e-07
## Residual       10.0    10.435533   1.043553       NaN           NaN
```

In jeder Stadt ist der Faktor `AUTO` hoch signifikant.

e) Wir wiederholen die Zweiweg-Faktoranalyse ohne die Stadt San Francisco: (zu **R**)

```
fit = ols("KMP4L ~ C(STADT,Sum)*C(AUTO,Sum)", data=df[df["STADT"]!="San
Francisco"]).fit()
anova_lm(fit)
```

```
##                               df       sum_sq  ...          F        PR(>F)
## C(STADT, Sum)                1.0     0.926853  ...   0.575105  4.570817e-01
## C(AUTO, Sum)                 4.0   349.513196  ...  54.217534  1.870272e-10
## C(STADT, Sum):C(AUTO, Sum)   4.0    11.066294  ...   1.716637  1.858312e-01
## Residual                    20.0    32.232487  ...        NaN           NaN
##
## [4 rows x 5 columns]
```

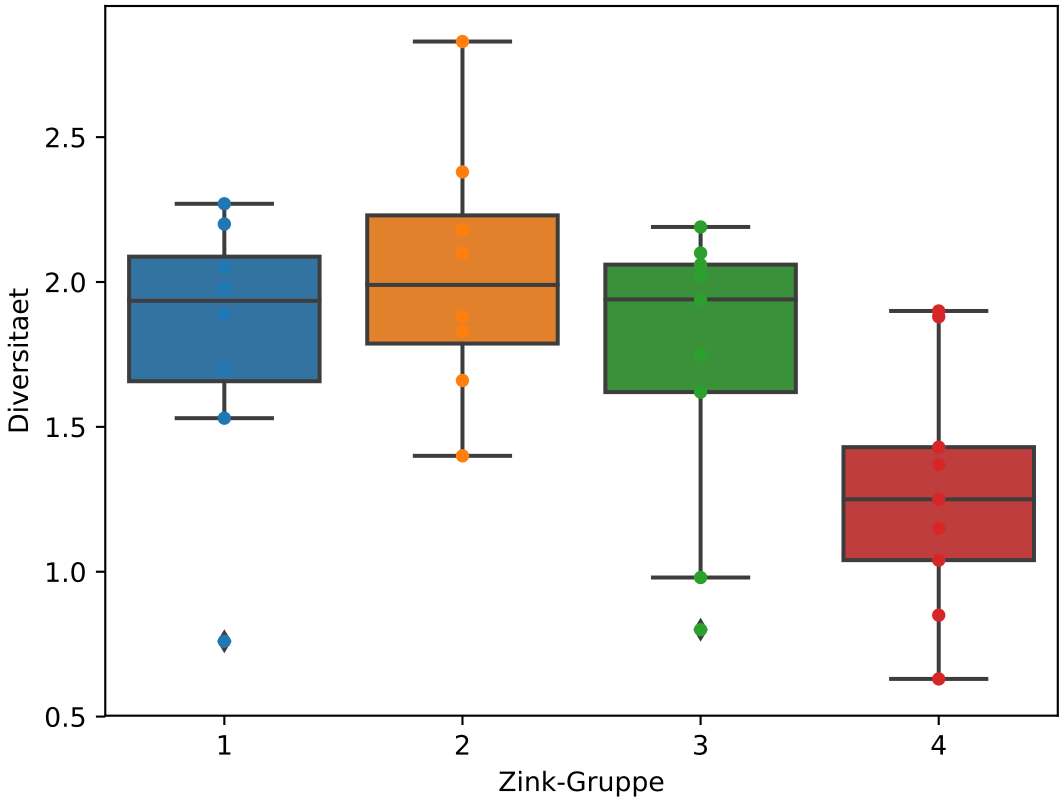
Weder die Wechselwirkung noch die Stadt sind auf dem 5 % Niveau signifikant. Dies ist aufgrund des Interaktionsplots auch nicht erstaunlich.

Lösung 10.5

a) (zu **R**)

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import seaborn as sns
stream = pd.read_csv('../../../Themen/Varianzanalyse/Uebungen_de/Daten/stre
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
sns.stripplot(x="ZNGROUP", y="DIVERSITY", data=stream)
plt.xlabel("Zink-Gruppe")
plt.ylabel("Diversitaet")
sns.boxplot(x="ZNGROUP", y="DIVERSITY", data=stream)
plt.xlabel("Zink-Gruppe")
```

```
plt.ylabel("Diversitaet")
```



Der Boxplot weist zwei Ausreisser auf. Im Stripchart scheinen diese allerdings keine Ausreisser mehr zu sein. Wenn nur wenige Datenpunkte vorhanden sind, so ist ein Boxplot nicht eine sehr vorteilhafte Visualisierungstechnik. Die Verteilung der Datenpunkte kann besser mit Stripcharts visualisiert werden. Zinkgruppe 4 weist eine signifikant tiefere Biodiversität auf.

b) (zu **R**)

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
from patsy.contrasts import Sum
stream = pd.read_csv('../../../Themen/Varianzanalyse/Uebungen_de/Daten/stream.dat',
sep=' ', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
fit = ols("DIVERSITY ~ C(ZNGROUP, Sum)", data=stream).fit()
print(anova_lm(fit))
```

```
##                        df     sum_sq   mean_sq        F    PR(>F)
## C(ZNGROUP, Sum)    3.0   2.566612  0.855537  3.93869  0.01756
## Residual          30.0   6.516411  0.217214      NaN       NaN
```

Das Gruppenmittelmodell lautet

$$Y_i = \mu + \alpha_i + \varepsilon_i$$

wobei μ den (globalen) Mittelwert α_i die Behandlungseffekte bezeichnen. Der F-Test ergibt ein signifikantes Resultat mit einem P-Wert von 0.018. Es gibt also einen signifikanten Unterschied der Biodiversität zwischen den unterschiedlichen Zinkgruppen.

c) (zu **R**)

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
from patsy.contrasts import Sum
stream = pd.read_csv('../../../Themen/Varianzanalyse/Uebungen_de/Daten/stream.dat', sep='
', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
fit = ols("DIVERSITY ~ C(ZNGROUP, Sum)", data=stream).fit()
print(fit.summary())
```

```
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:              DIVERSITY   R-squared:                       0.283
## Model:                            OLS   Adj. R-squared:                  0.211
## Method:                 Least Squares   F-statistic:                     3.939
## Date:                Tue, 28 Apr 2020   Prob (F-statistic):             0.0176
## Time:                        01:03:00   Log-Likelihood:                -20.159
## No. Observations:                  34   AIC:                             48.32
## Df Residuals:                      30   BIC:                             54.42
## Df Model:                           3
## Covariance Type:            nonrobust
## ========================================================================================
##                            coef    std err          t      P>|t|      [0.025      0.975]
## ----------------------------------------------------------------------------------------
## Intercept                1.7064      0.080     21.312      0.000       1.543       1.870
## C(ZNGROUP, Sum)[S.1]     0.0911      0.141      0.644      0.524      -0.198       0.380
## C(ZNGROUP, Sum)[S.2]     0.3261      0.141      2.307      0.028       0.037       0.615
## C(ZNGROUP, Sum)[S.3]     0.0114      0.136      0.084      0.934      -0.266       0.289
## ==============================================================================
## Omnibus:                        2.067   Durbin-Watson:                   1.196
## Prob(Omnibus):                  0.356   Jarque-Bera (JB):                1.746
## Skew:                          -0.542   Prob(JB):                        0.418
## Kurtosis:                       2.757   Cond. No.                         2.12
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Das Gruppenmittelmodell lautet

$$Y_i = \mu + \alpha_i + \varepsilon_i$$

Somit sind $\hat{\mu} = 1.7064$, $\hat{\alpha}_1 = 0.0911$, und somit ist $\hat{\mu}_1 = 1.7064 + 0.0911 = 1.7975$, $\hat{\mu}_2 = 1.7064 + 0.3261 = 2.0325$, $\hat{\mu}_3 = 1.7975 + 0.0114 = 1.8089$ und $\hat{\mu}_4 = 1.7975 - 0.4286 = 1.3689$.

d) (zu **R**)

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
from patsy.contrasts import Sum
stream = pd.read_csv('../../../Themen/Varianzanalyse/Uebungen_de/Daten/stre
sep=' ', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
fit2 = ols("DIVERSITY ~ C(ZNGROUP, Sum) + C(STREAM, Sum)", data=stream).fit
print(anova_lm(fit2))
```

```
##                       df     sum_sq   mean_sq         F    PR(>F)
## C(ZNGROUP, Sum)      3.0   2.566612  0.855537  6.660454  0.001852
## C(STREAM, Sum)       5.0   3.305153  0.661031  5.146196  0.002216
## Residual            25.0   3.211258  0.128450       NaN       NaN
```

Die Variable **STREAM** ist signifikant, man wird sie als Block-Variable interpretieren, da sie an und für sich für die Fragestellung nicht von Bedeutung ist. Dennoch ist sie signifikant und sollte in der Analyse berücksichtigt werden. Eine Interaktions-Analyse zwischen **STREAM** und **ZNGROUP** ergibt, dass es keine Interaktion zwischen den beiden Variablen gibt.

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
from patsy.contrasts import Sum
stream = pd.read_csv('../../../Themen/Varianzanalyse/Uebungen_de/Daten/stream.dat', sep='
', header=0)
stream["ZNGROUP"] = stream["ZNGROUP"].apply(str)
fit2 = ols("DIVERSITY ~ C(ZNGROUP, Sum) * C(STREAM, Sum)", data=stream).fit()
print(anova_lm(fit2))
```

```
##                                     df    sum_sq   mean_sq         F    PR(>F)
## C(ZNGROUP, Sum)                    3.0  2.566612  0.855537  7.118441  0.002642
## C(STREAM, Sum)                     5.0  3.305153  0.661031  5.500059  0.003430
## C(ZNGROUP, Sum):C(STREAM, Sum)    15.0  2.211068  0.147405  1.226469  0.340100
## Residual                          17.0  2.043163  0.120186       NaN       NaN
```

R-Code