Bericht über die Datenanalyse des SHP-Datensatzes

Dimitriev, Erismann, Hosseini

2022-04-28

Dokumentation Projekt 2

Einleitung

Dieser Bericht behandelt folgende Fragestellung:

"Welche Faktoren stehen mit einer Veränderung der allgemeinen Lebenszufriedenheit seit Beginn der Verbreitung von Covid-19 in Beziehung?"

Die Zielgrösse '*Änderung der allgemeinen Lebenszufriedenheit von 2019 auf 2020'* soll untersucht werden. Dazu kommen Variablen, die mögliche Einflussfaktoren darstellen. Des weiteren bestimmten wir drei zusätzliche Variablen, die möglicherweise einen Einfluss auf die Zielvariable haben. Die verwendeten Daten stammen vom Swiss Household Panel (SHP). Es sind nur volljährige Personen berücksichtigt worden.

Organisation

Woche	Auftrag	Zuständigkeit
28.03 - 01.04	• Einarbeiten in den Auftrag	Team
	• Datenstruktur planen und festlegen: git.zhaw	Patrick
	• Daten herunterladen	Ahmad
	• Einlesen der ersten Daten in R	Patrick
	• Uns mit den erhaltenen Daten vertraut machen	Team
	• Definieren der eigenen Variablen	Team
	• Erstellen der Dokumentation	Team
04.04 - 08.04	• Alle Daten einlesen	Patrick
	• Daten gemäss Vorgaben aufbereiten	Ahmad
	• Vorgegebene Variablen einlesen und wenn nötig definieren	Team
	• Eigene Variablen erstellen	Patrick
11.04 - 15.04	• Re-Check der definierten Variablen	Team
	• Prüfen wo es NA's hat, nötigenfalls Annnahmen treffen	Patrick
	• Abgleich der übriggebliebenen Daten mit Daten vom BFS	Stojche

Woche	Auftrag	Zuständigkeit
18.04 - 22.04	• EXPD • Auswertung der Daten • R-Markdown erstellen für Abgabe	Ahmad Stojche Patrick

Variablen

Vordefinierte Variablen:

1 Änderung der allgemeinen Lebenszufriedenheit von 2019 auf 2020 (schlechter, gleich, besser)
2 Geschlecht
3 Alter
4 Kanton
5 Gemeindetyp
6 Höchstes Ausbildungszertifikat
7 Beziehungsstatus (in Beziehung, nicht in Beziehung)
8 Änderung des Beziehungsstatus von 2019 auf 2020 (Trennung, keine Änderung, Neu in Beziehung)
9 Kurzfristige Änderung der Arbeitssituation nach Pandemiebeginn (ja = wurde Arbeitslos, Kurzarbeit, arbeitet weniger z.B. wegen Kinderbetreuung, eigenes Geschäft ist direkt von Pandemie betroffen, sonst nein)

10 Corona Infektion (nein, ja)
11 Haushaltstyp
12 Jährliches Netto-Haushaltseinkommen pro Haushaltsmitglied
13 Finanzielle Probleme in der Jugend
14 War Person mindestens einmal Arbeitslos zwischen 2015 und 2019?

Selbstdefinierte Einflussfaktoren

1. Illness / Accident (P20L01)
   Hypothese: In der Pandemie kam es zu weniger Unfällen, wegen den strikten Einschränkungen in der Mobilität, deshalb sind die Menschen glücklicher.

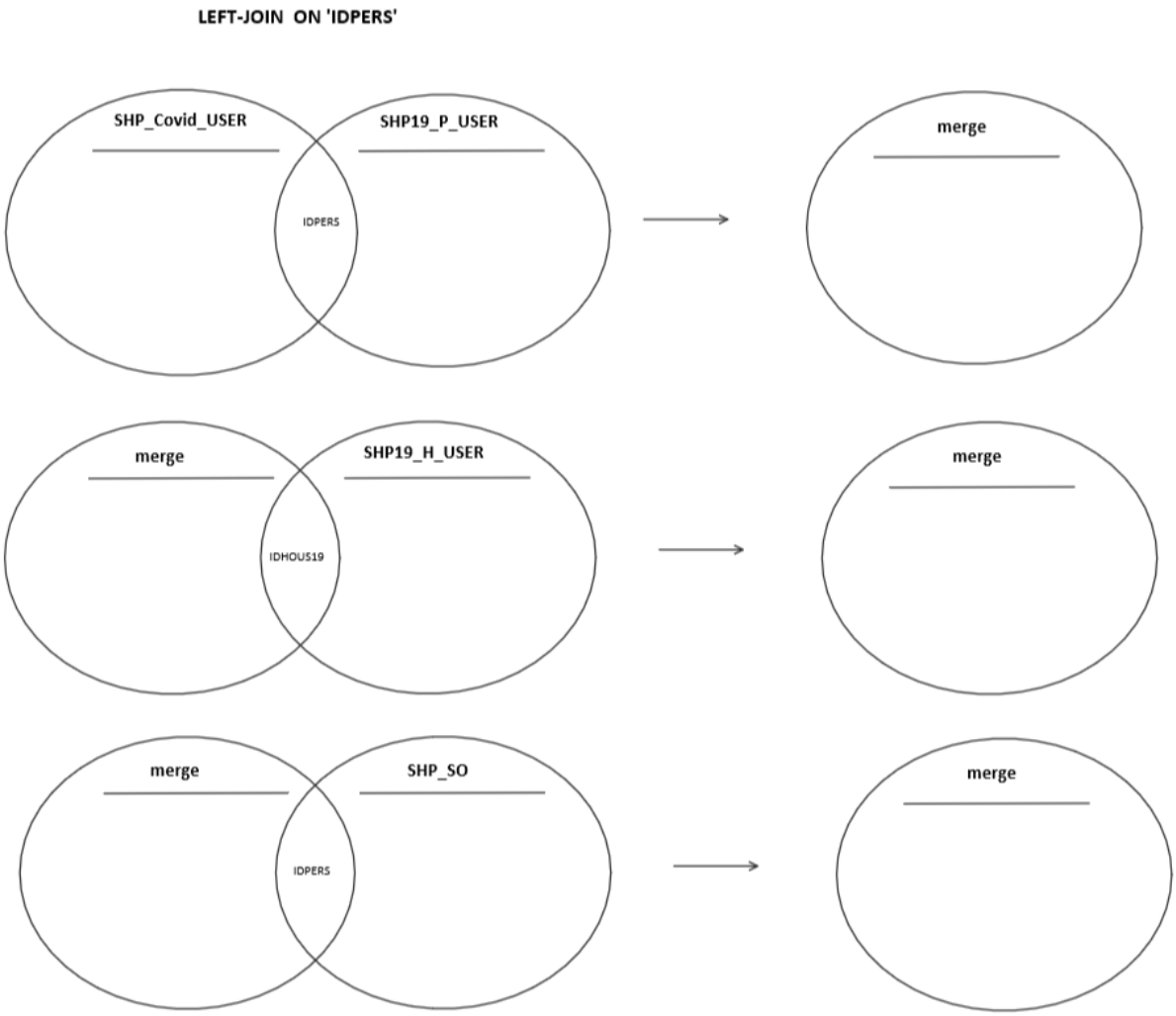
2. Happy with Partner(P20F54)
   Hypothese: Zielgrösse ist abhängig von der Zufriedenheit mit dem Partner.

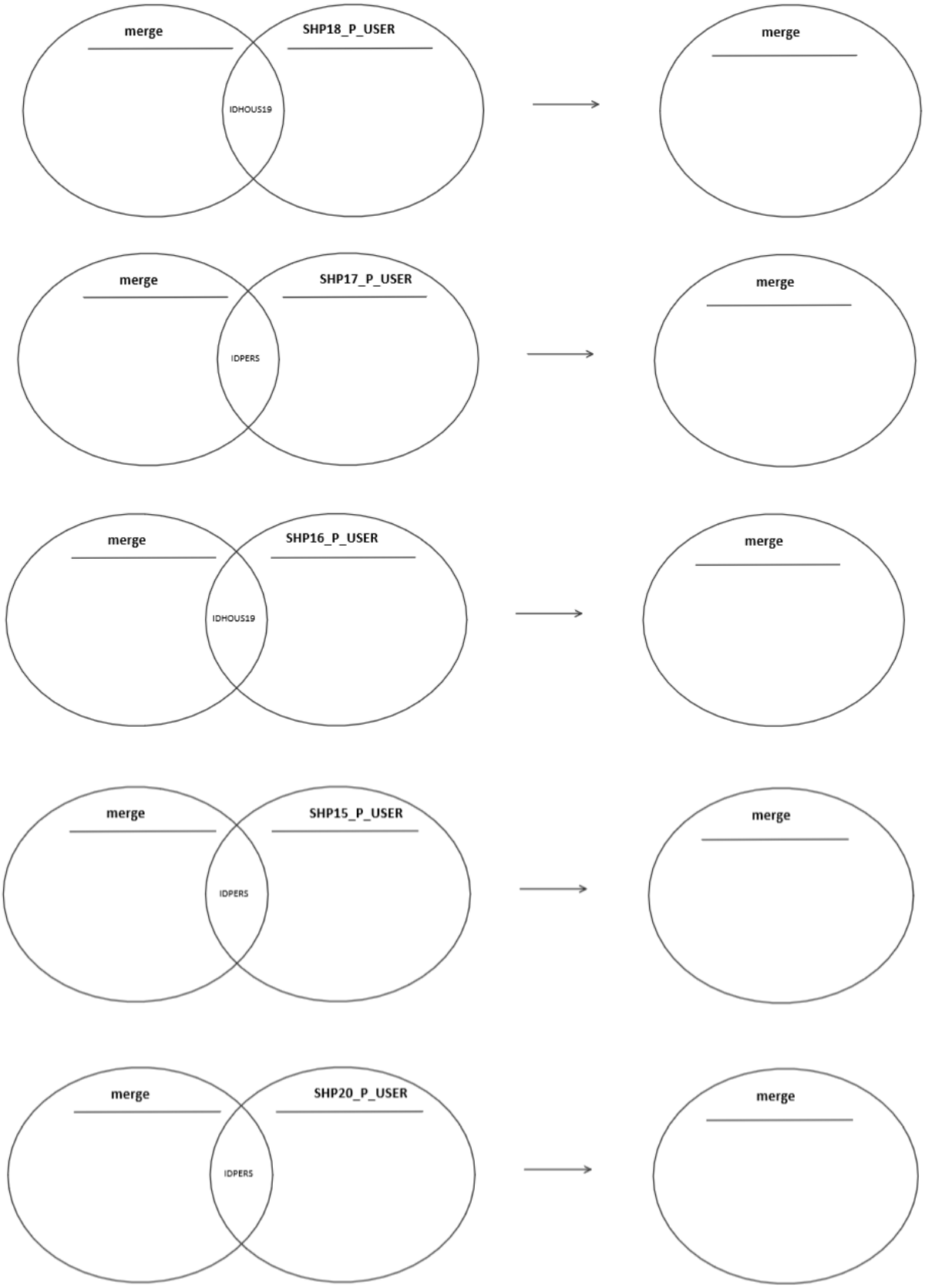
3. Trust in Federal Government(P17P04)
   Hypthese: Wer durch die Pandemie das Vertrauen in die Regierung verlor, ist unglücklicher.
   Die Variable ist den Jahren 2019 und 2018 nicht verfügbar, deshalb griffen wir auf das Jahr 2017 zurück.

Datenaufbereitung

Die Datenaufbereitung erwies sich als keine leichte Kost. Zuerst selektierten wir die benötigten Variablen, um die Zusammenführung der Daten nicht unnötig gross zu gestalten. Danach wurden die Datensätze wie folgt zusammengeführt (Alle sind mit Left-Joins zusammengeführt).

Zusammenführen der Daten gemäss Skizze




merge
SHP18_P_USER
IDHOUS19
merge
merge
SHP17_P_USER
IDPERS
merge
merge
SHP16_P_USER
IDHOUS19
merge
merge
SHP15_P_USER
IDPERS
merge
merge
SHP20_P_USER
IDPERS
merge

Unser Datensatz 'merge', beinhaltet somit alle benötigten Variablen, um die Bereinung der Daten zu starten.

Resultate des Parallel-Codings

Stojche und Ahmed führten gleichzeitig eine explorative Datenanalyse der Daten durch und verglichen schliesslich die Analyse. Ahmad erstellte Diagramme, die sich auf die Verteilung der einzelnen Kategorien konzentrieren. Stojche erstellte Diagramme, die sich sowohl auf die Verteilung jeder Kategorie als auch zusätzlich auf die Anzahl der Datenpunkte, die zu dieser Kategorie gehören, konzentrieren. Seine Variante deutet darauf hin, ob die Verteilung in jeder Kategorie im Verhältnis zur Anzahl der Datenpunkte in dieser Kategorie sinnvoll ist.

Datenqualität

NA-Werte:

Variable	Anzahl (% - Anteil der Gesamtdaten)	Grund
Change_satisfaction	63 (~1%)	Der Grossteil der NA's ist aus des Covid-Daten(C20PC44), es handelt sich unserer Meinung nach um eine geringe Anzahl NA' bezogen auf die Gesamtdaten.
Relation_status	72 (~1,5%)	Coviddatensatz beinhaltet fehlende Daten.
Change_in_relation	170 (~3,5%)	Die Variable (C20D29) beinhaltet 72 NA's, die sich durch die neue Definition der Variable auf 170 erhöhten.
Workingsituation	151 (~3%)	Entstehung der NA's durch die Kombination verschiedener Variablen. Mögliche Gründe für diese NA's könnten sein, dass Leute ihre Arbeitssituation preisgeben wollen.
Corona_infection	99 (~2%)	Fehlende Werte sind gering.
Household_type	3 (~0.05%)	Werte kommen von der Variable.
Income	383 (~7%)	Grosse Anzahl fehlende Werte. NA's sind bereits von der Variable (I19HTYN). Es könnte durch Schweizer-Spiessbürgertum erklärt werden. Man redet nicht über Geld.
Financial_problem	309 (~6%)	Grosse Anzahl fehlende Werte. Variable ist keine Kombination aus mehreren Variablen! Man sieht ähnlich Situation wie Income.
Illness_accident	368 (~6.5%)	Die allermeisten Werte entstammen der Variable (P20L01), die Variable (P19L01) enthält nur drei NA's.
Happy_with_partner	1585 (~28%)	Die beiden Grundvariablen (P19F54,P20F54) beinhalten beide über 1000 fehlende Werte.

Variable	Anzahl (% - Anteil der Gesamtdaten)	Grund
Trust_in_government	819 (~14%)	Die Variable (P17P04) enthält 533 fehlende Werte und die Variable (P20P04) enthält 390 NA's.
Total	4022 (~71%)	Wenn wir alle NA's entfernen würden, gingen drei-viertel der Beobachtungen verloren.

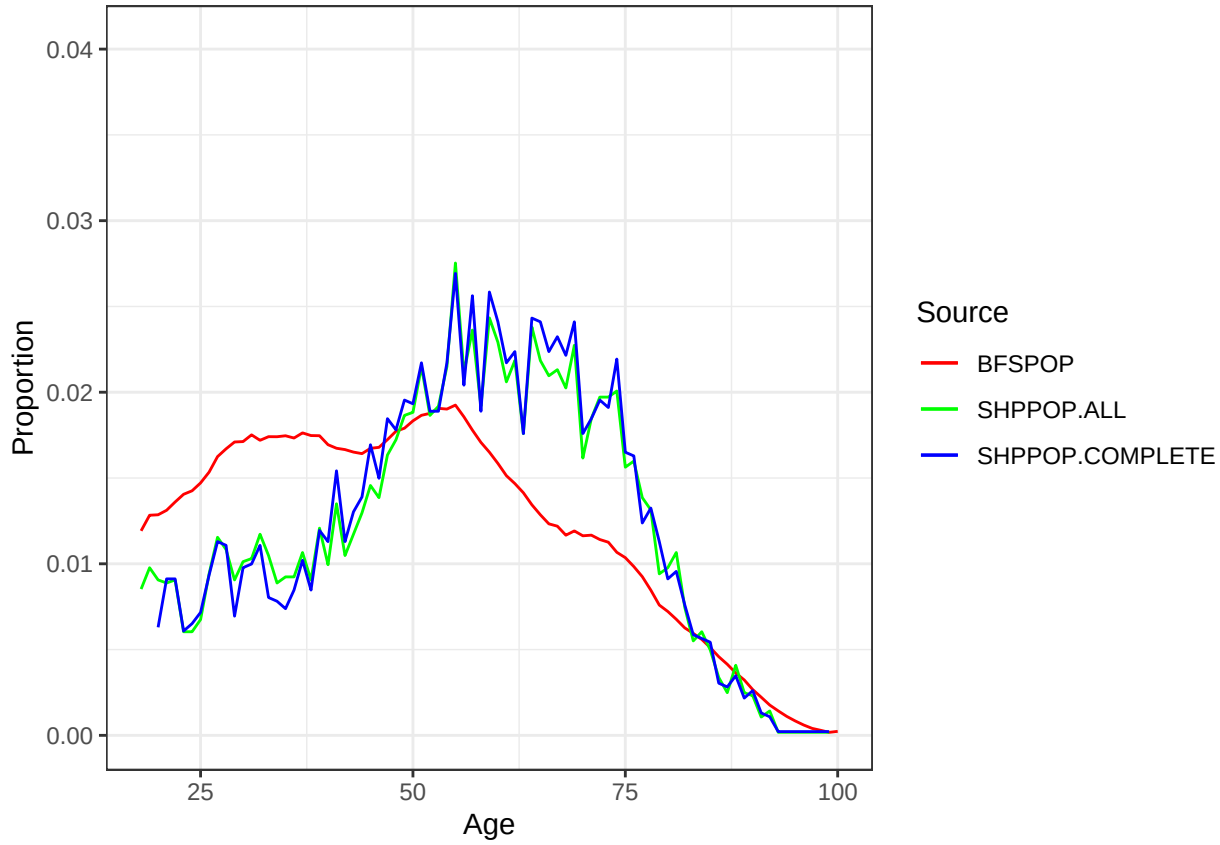
Fazit:

Variablen, die im Bezug auf die Gesamtbeobachtungen über 10% fehlende Werte haben, sollten für die Auswertung nicht verwendet werden, da die Stichprobe stark an ihrer Representativität einbüsst, wenn alle fehlenden Werte entfernt werden würden.

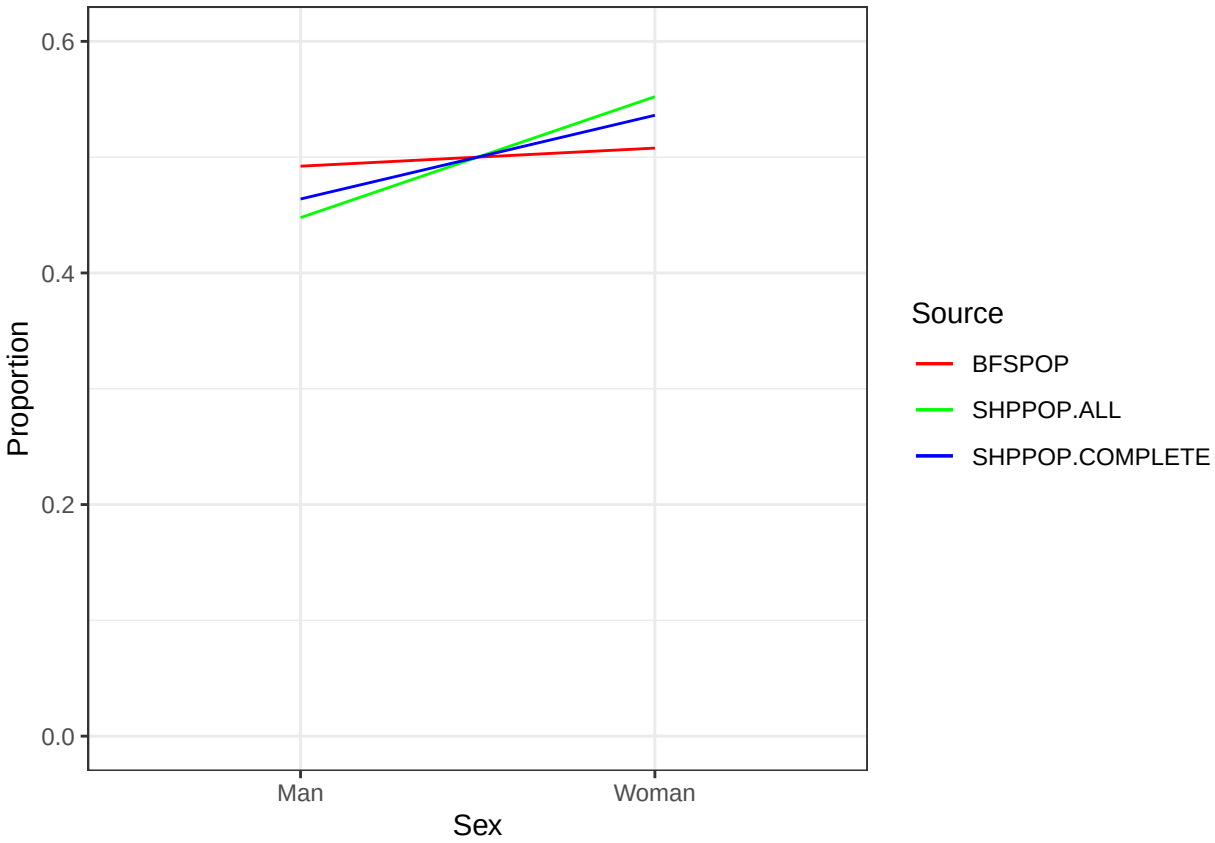
Deshalb änderten wir bei unseren selbstdefinierten Variablen alle NA's zum Zustand 'unknown', damit keine unnötigen Verluste der Beobachtungen in Kauf genommen werden müssen. Dadurch konnten wir die Verluste der NA's auf 1025 Beobachtungen reduzieren.

Plausibilität der Daten

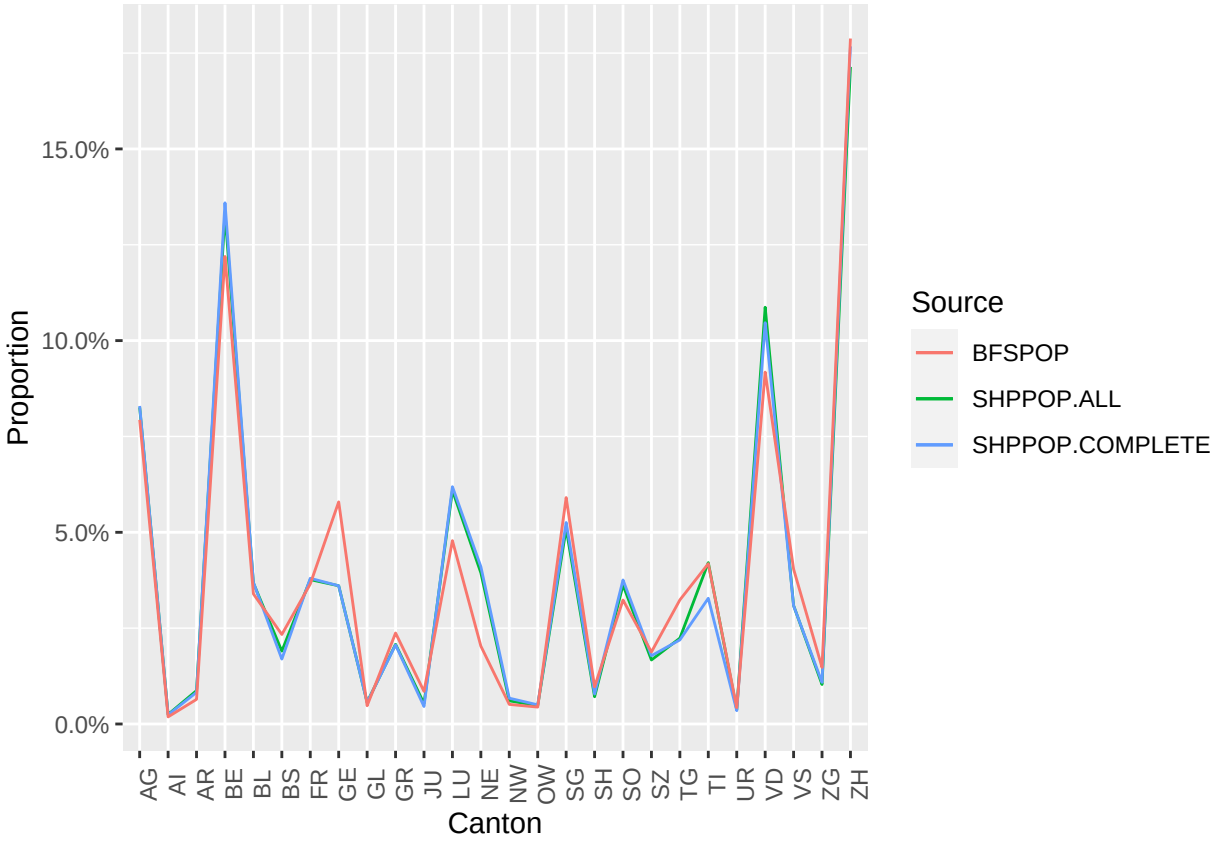
Um zu prüfen, dass die Stichprobe die Bevölkerung der Schweiz repräsentiert, vergleichen wir die Stichprobe mit Daten des BFS. Die drei Grafiken sollen darüber aufschliessen, wie repräsentativ unser Datensatz ist:



Wie sich auf der Grafik erkennen lässt, sind jüngere Altersgruppen, sprich die 25 bis 35-Jährigen Menschen in der Stichprobe unterrepräsentiert. Hingegen sind ältere Altersgruppen überrepräsentiert.



Im Bezug auf die Geschlechterverteilung der Stichprobe sind keine nennenswerte Unterschiede zu erläutern.

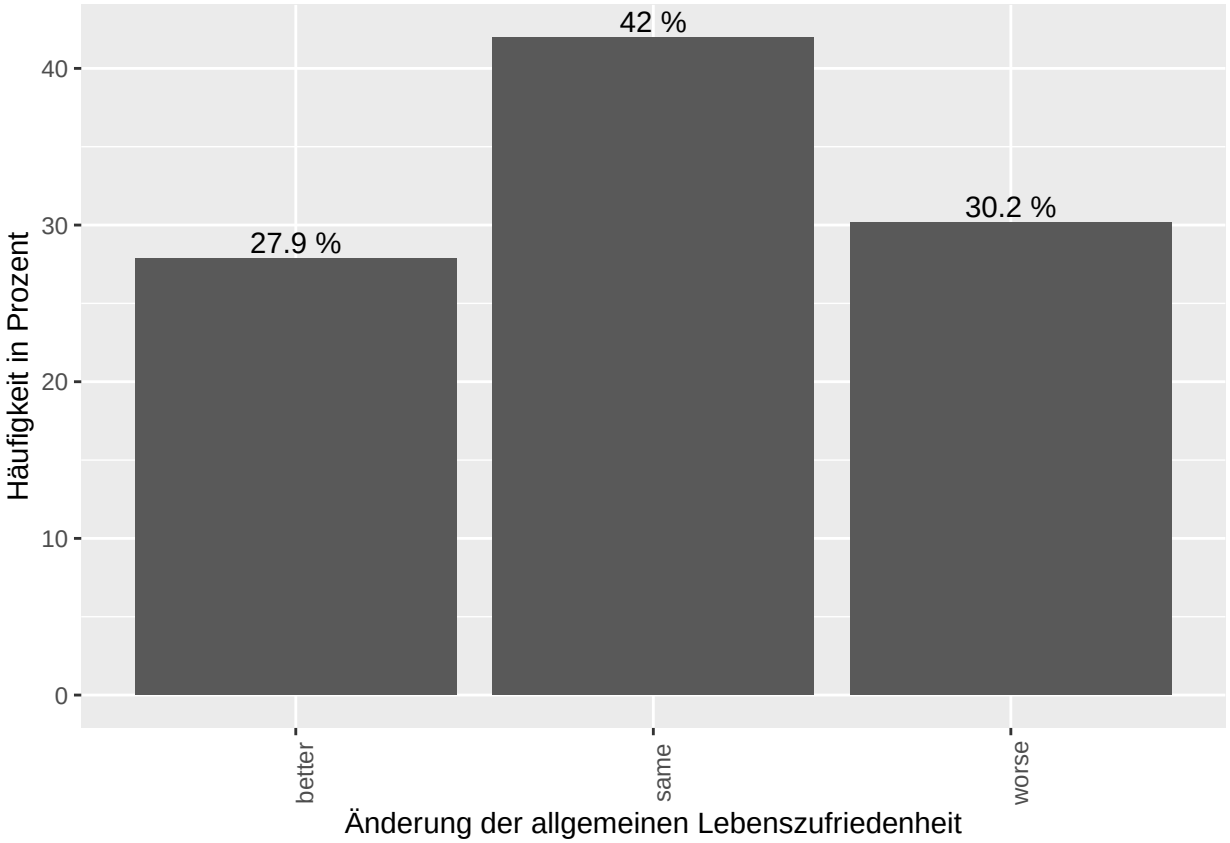


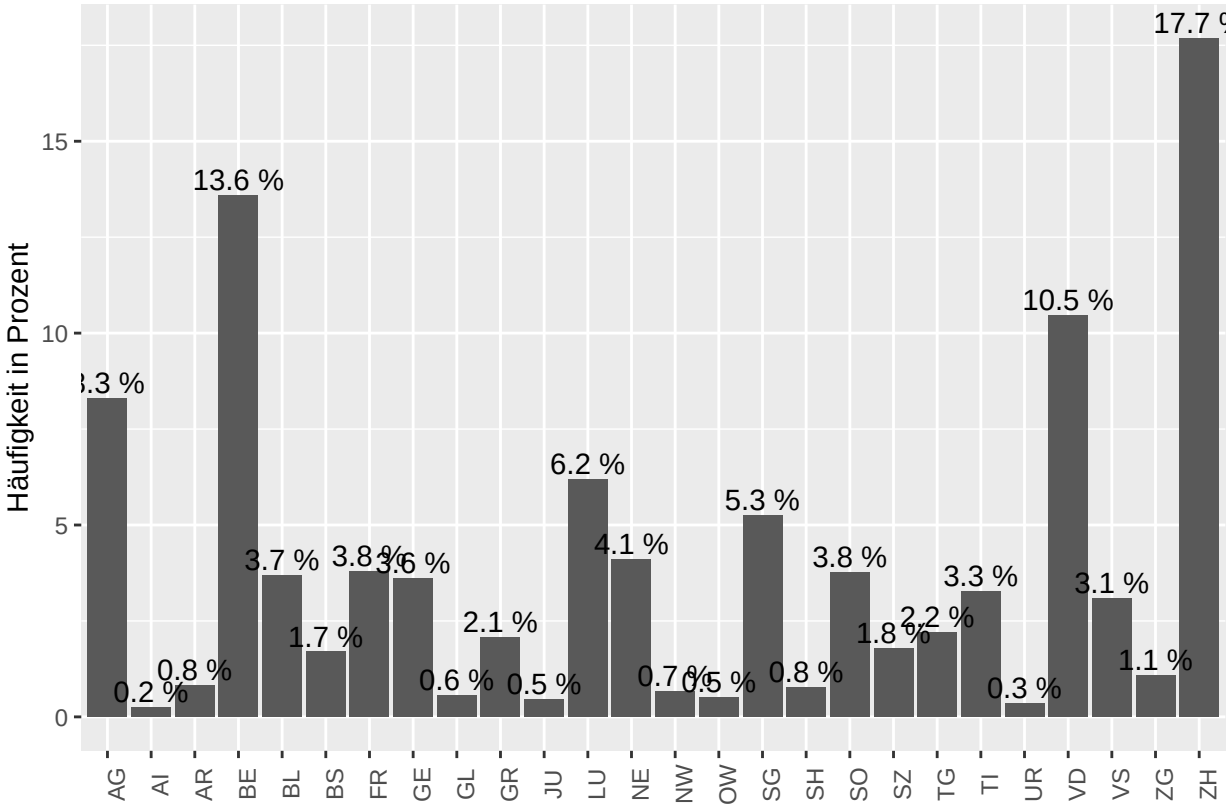
Die Vertretung der Kantone in der Stichprobe ist für die Schweizerbevölkerung representätiv.

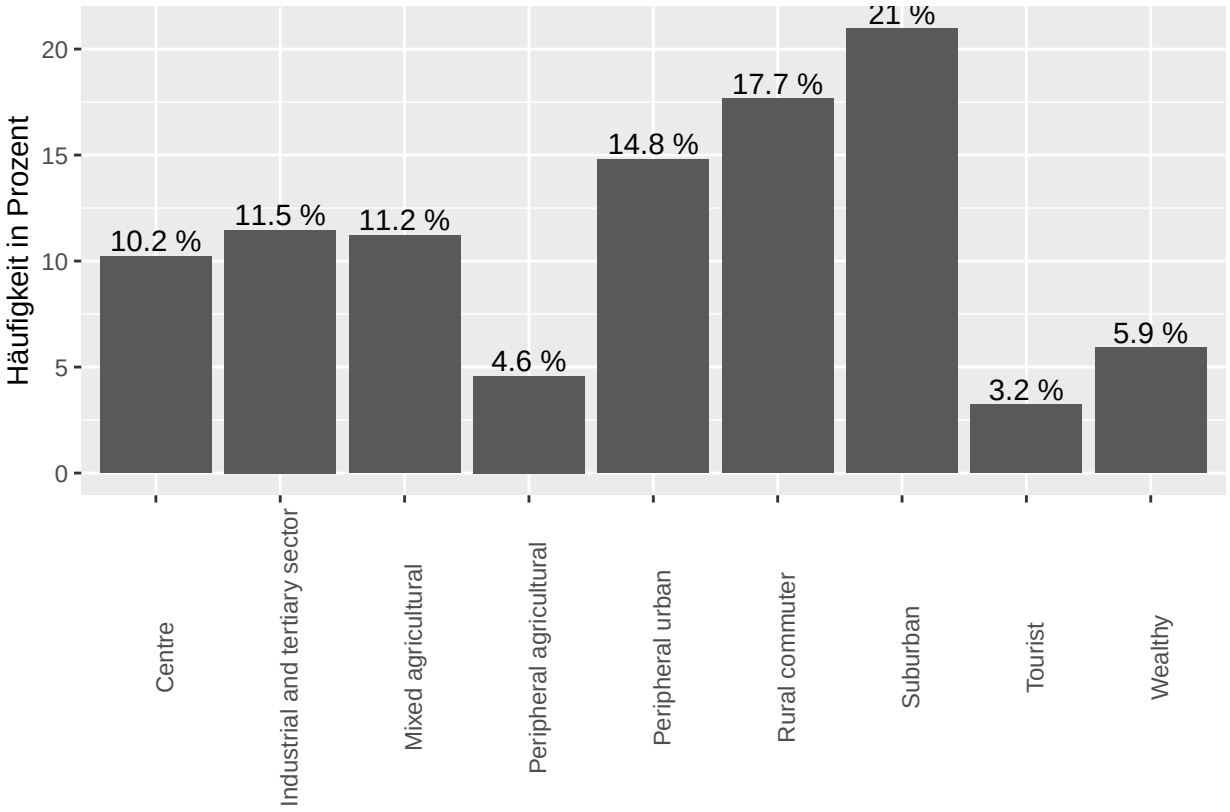
Auswertung

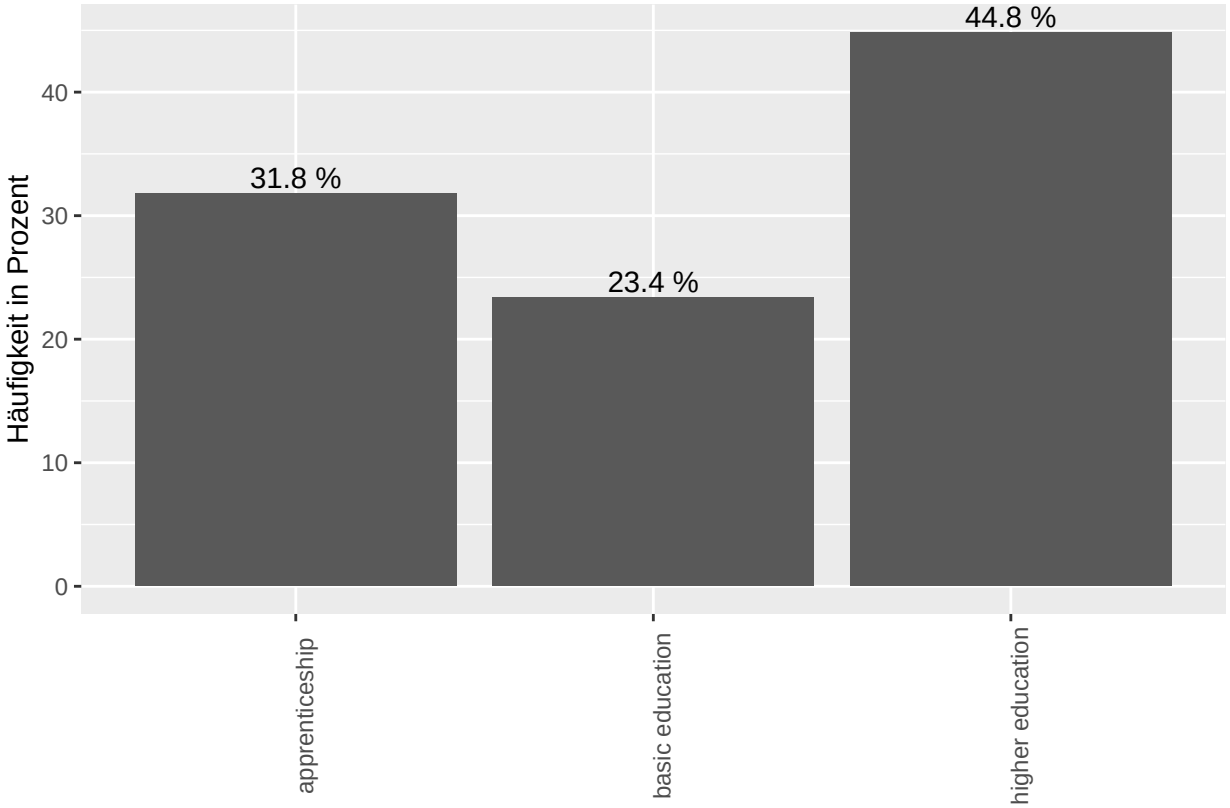
Univariate Grafiken

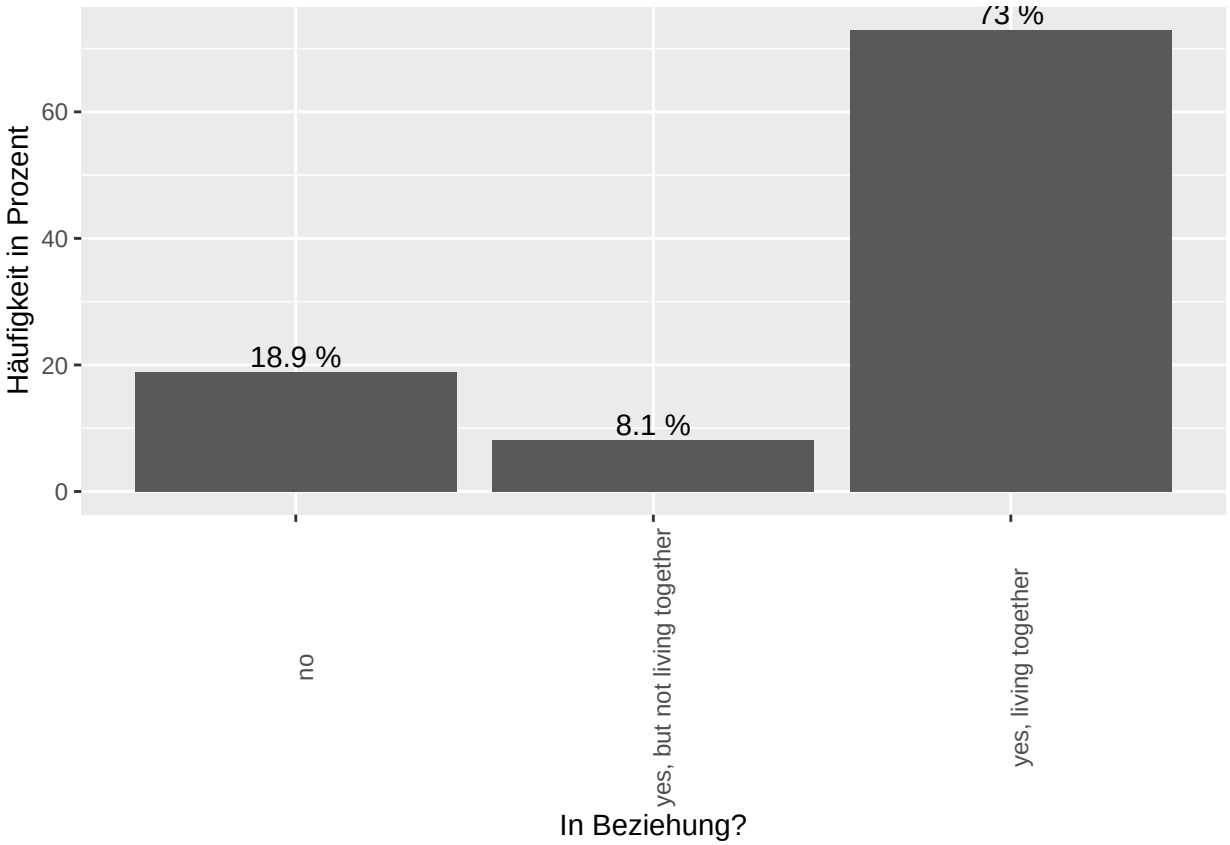
Die folgenden Abbildungen zeigen die explorative Datenanlyse mit jeweils einer Variable.

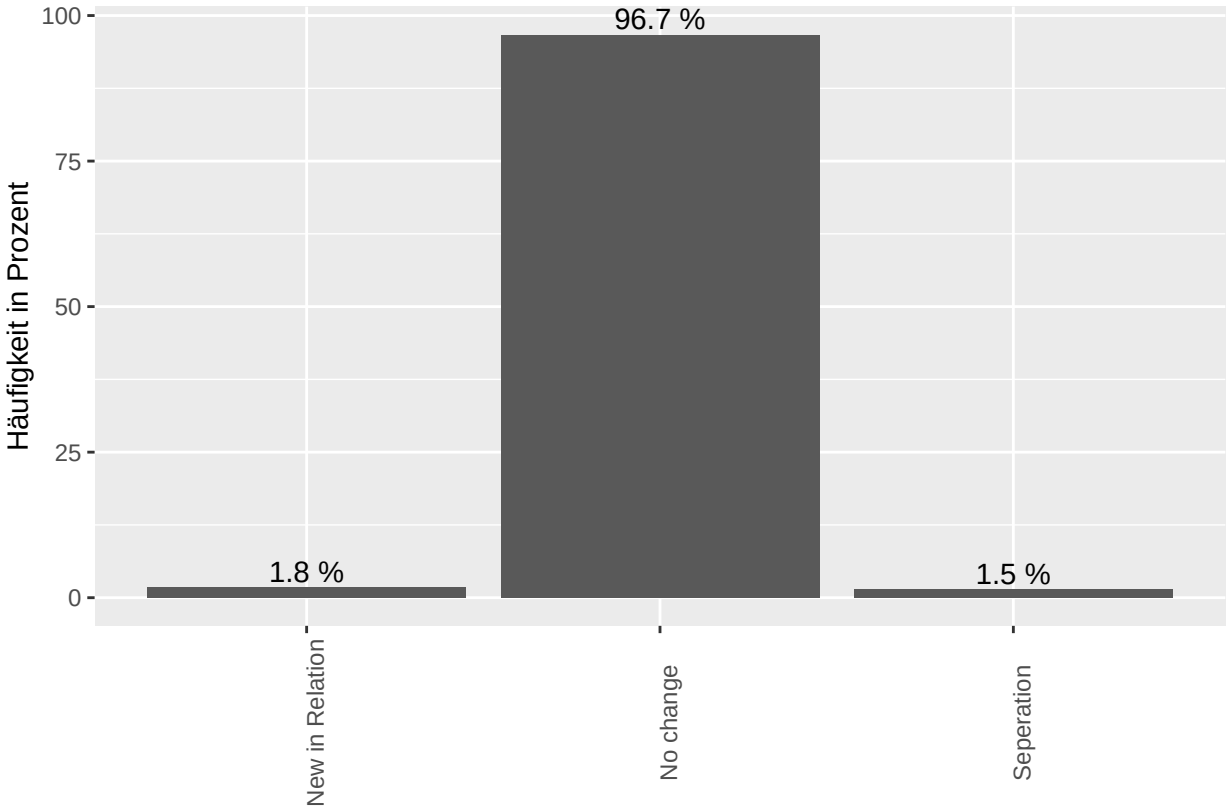


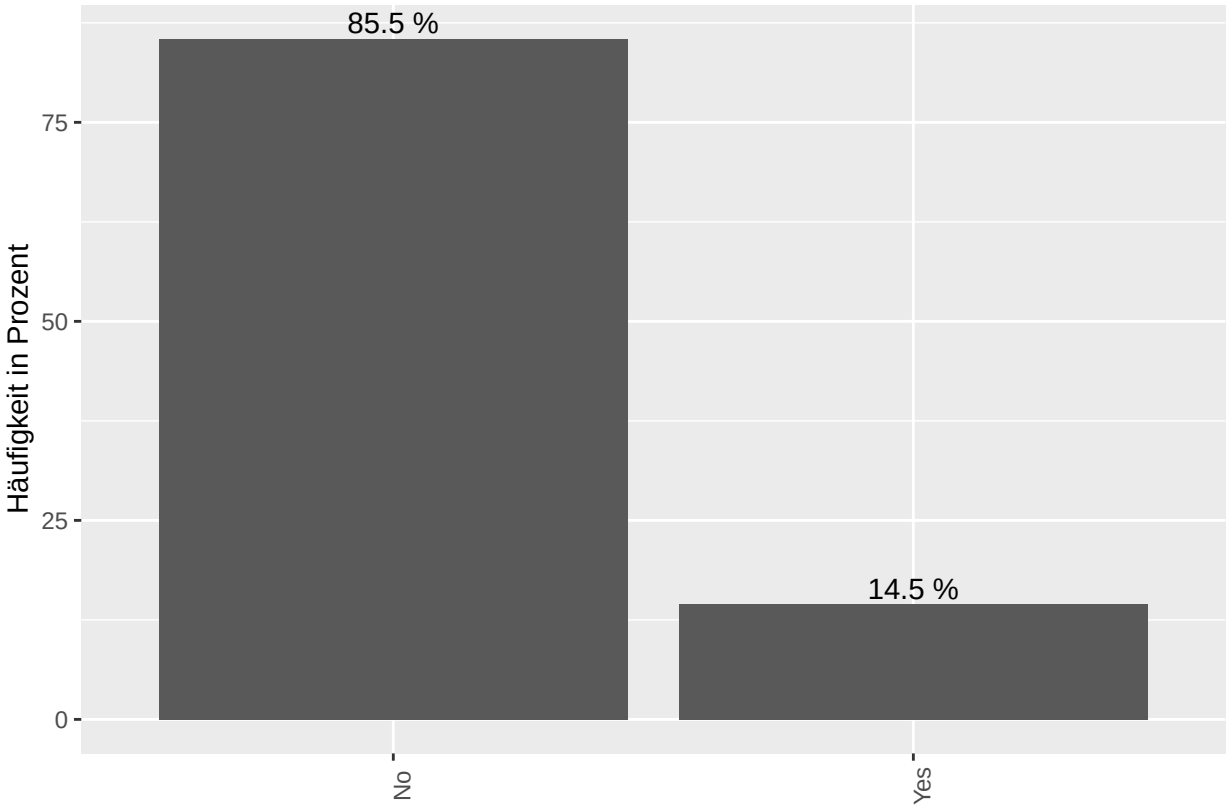

Häufigkeit in Prozent
0
5
10
15
3.3 %
0.2 %
0.8 %
13.6 %
3.7 %
1.7 %
3.8 %
3.6 %
0.6 %
2.1 %
0.5 %
6.2 %
4.1 %
0.7 %
0.5 %
5.3 %
0.8 %
3.8 %
1.8 %
2.2 %
3.3 %
0.3 %
10.5 %
3.1 %
1.1 %
17.7 %
AG
AI
AR
BE
BL
BS
FR
GE
GL
GR
JU
LU
NE
NW
OW
SG
SH
SO
SZ
TG
TI
UR
VD
VS
ZG
ZH

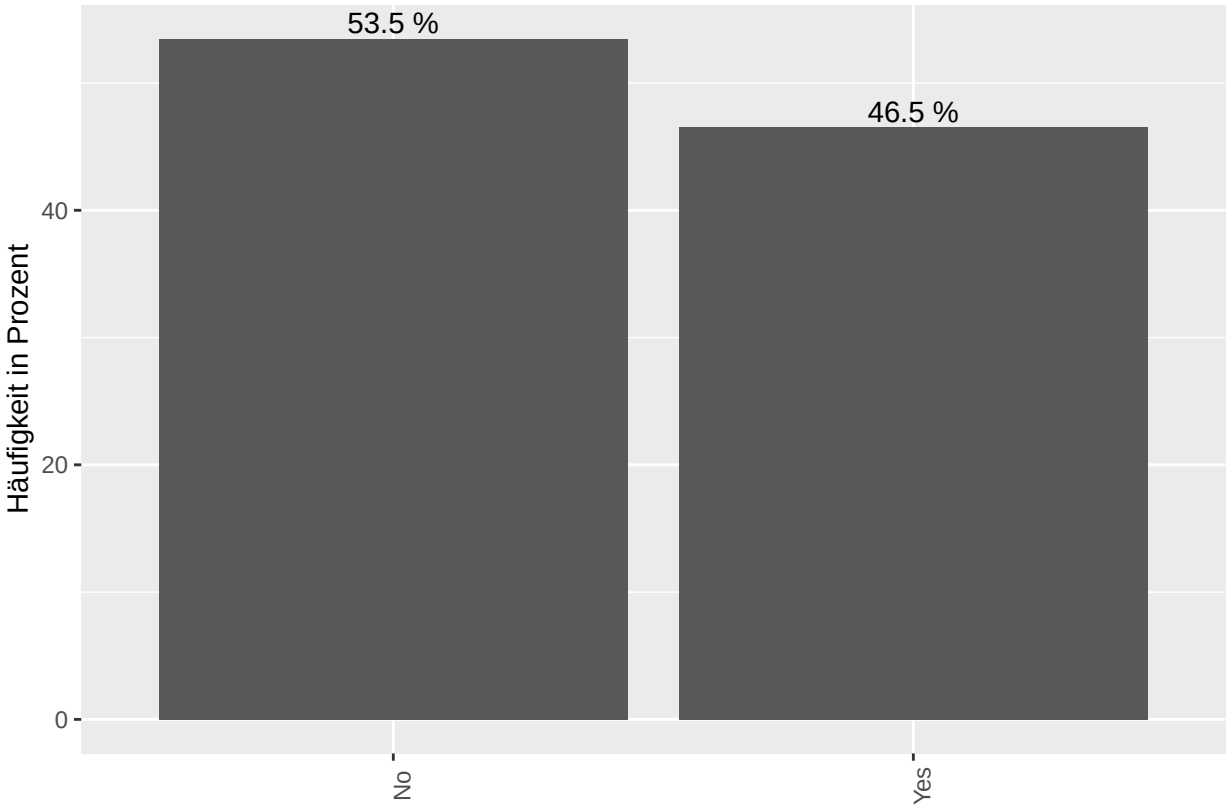

Häufigkeit in Prozent
0
5
10
15
20
10.2 %
11.5 %
11.2 %
4.6 %
14.8 %
17.7 %
21 %
3.2 %
5.9 %
Centre
Industrial and tertiary sector
Mixed agricultural
Peripheral agricultural
Peripheral urban
Rural commuter
Suburban
Tourist
Wealthy

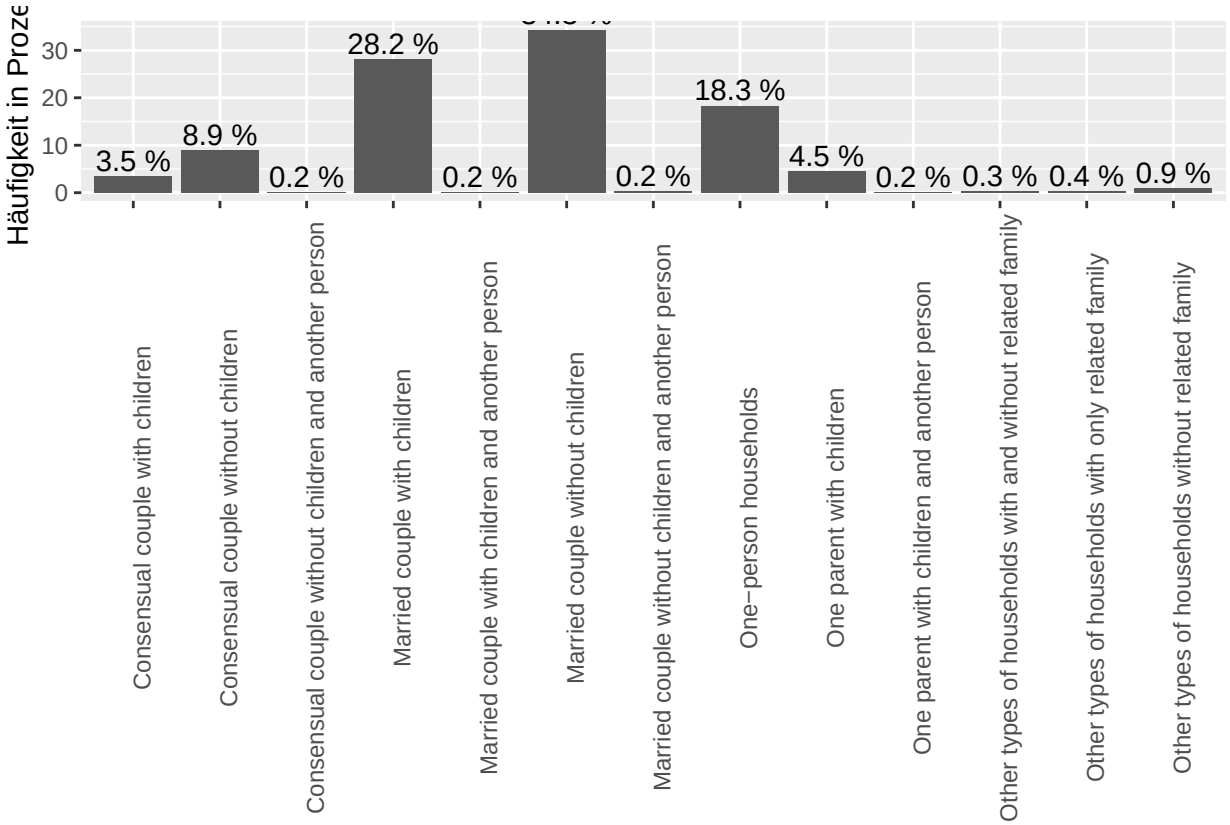

Häufigkeit in Prozent
0
10
20
30
40
31.8 %
23.4 %
44.8 %
apprenticeship
basic education
higher education

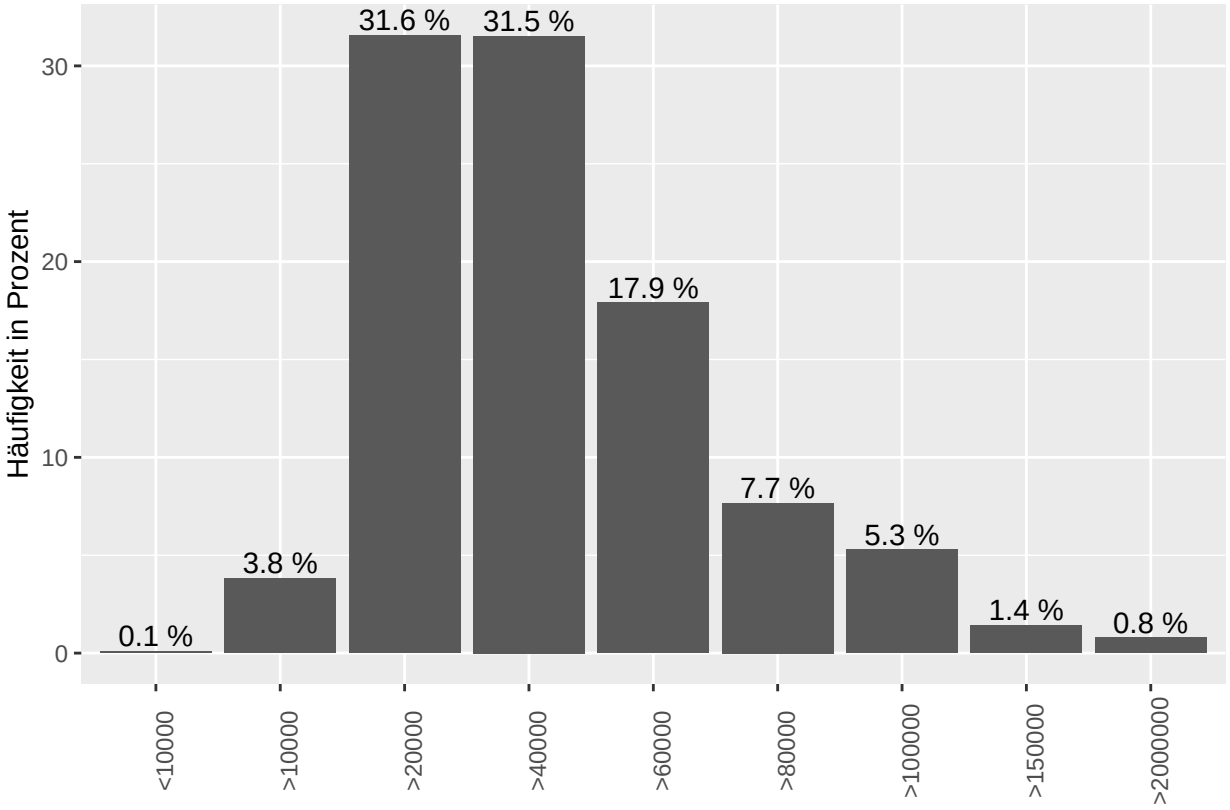

73 %
60
40
20
0
Häufigkeit in Prozent
18.9 %
8.1 %
no
yes, but not living together
yes, living together
In Beziehung?

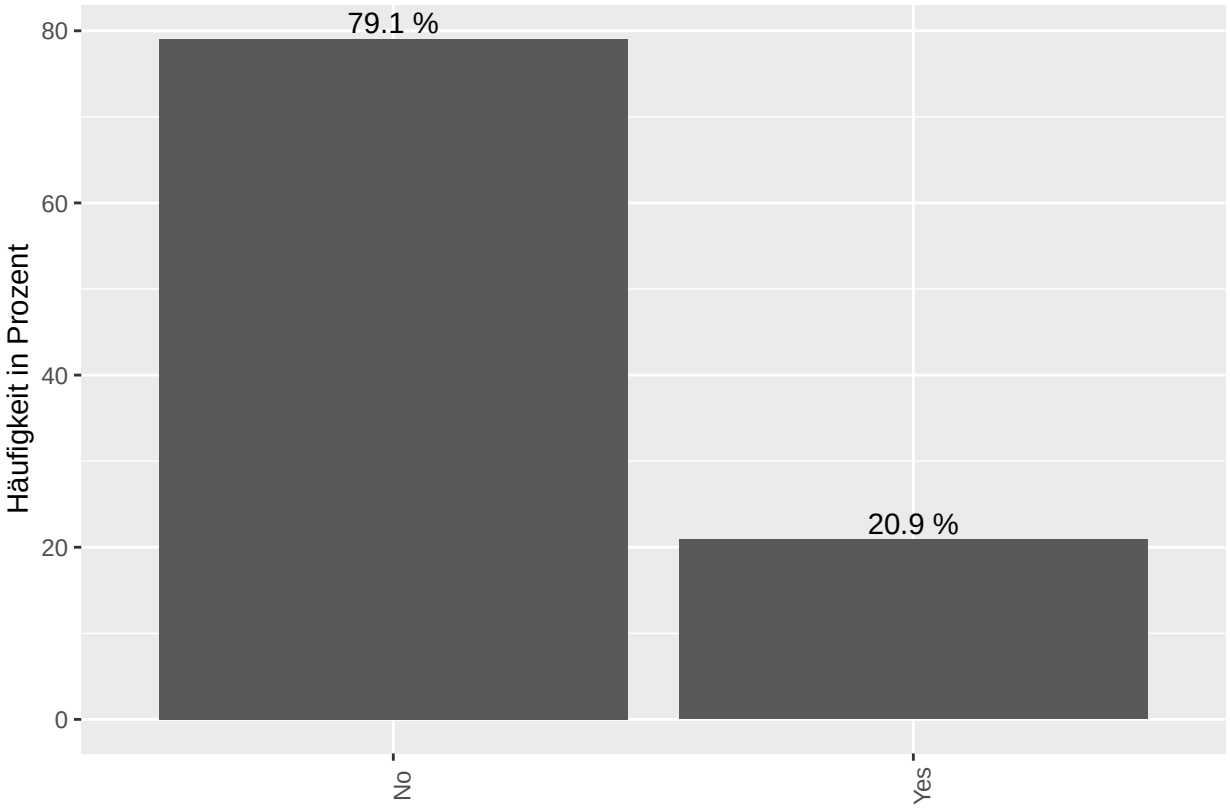

100
75
50
25
0
Häufigkeit in Prozent
96.7 %
1.8 %
1.5 %
New in Relation
No change
Seperation

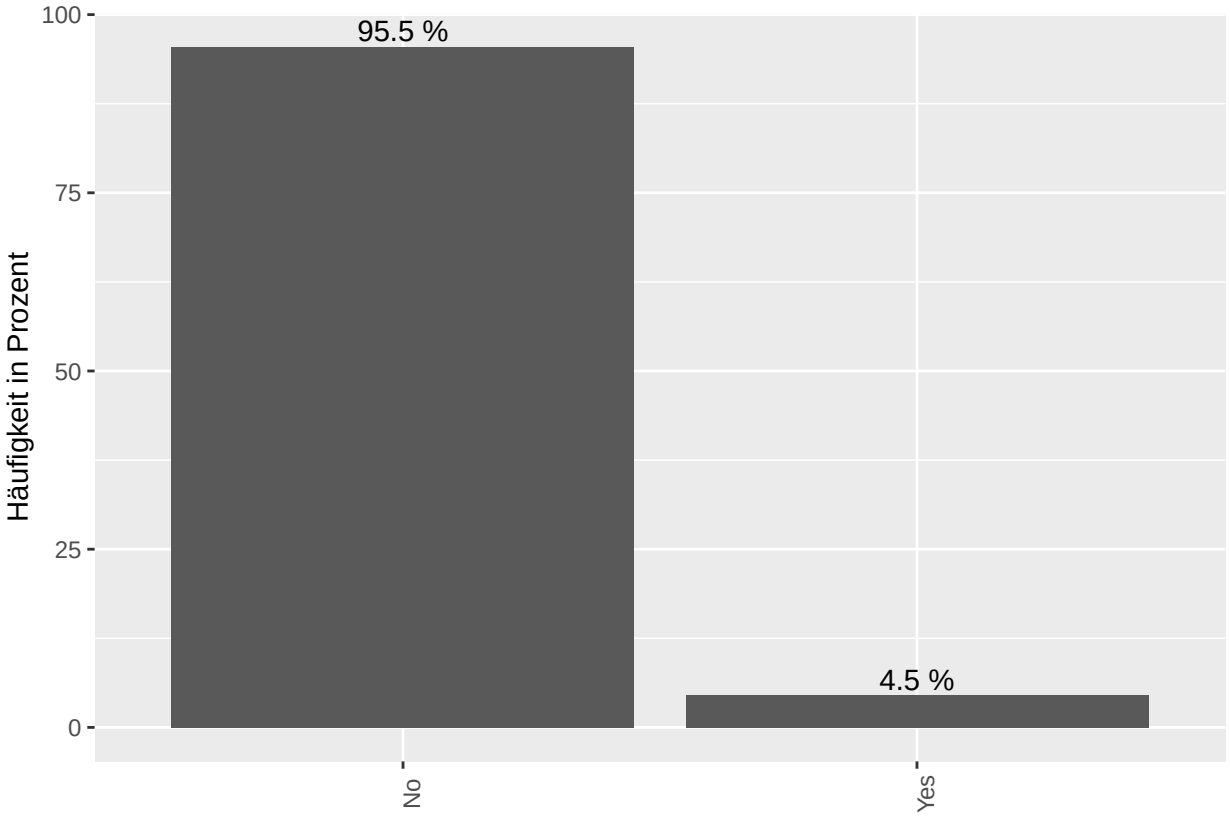

85.5 %
14.5 %
Häufigkeit in Prozent
75
50
25
0
No
Yes

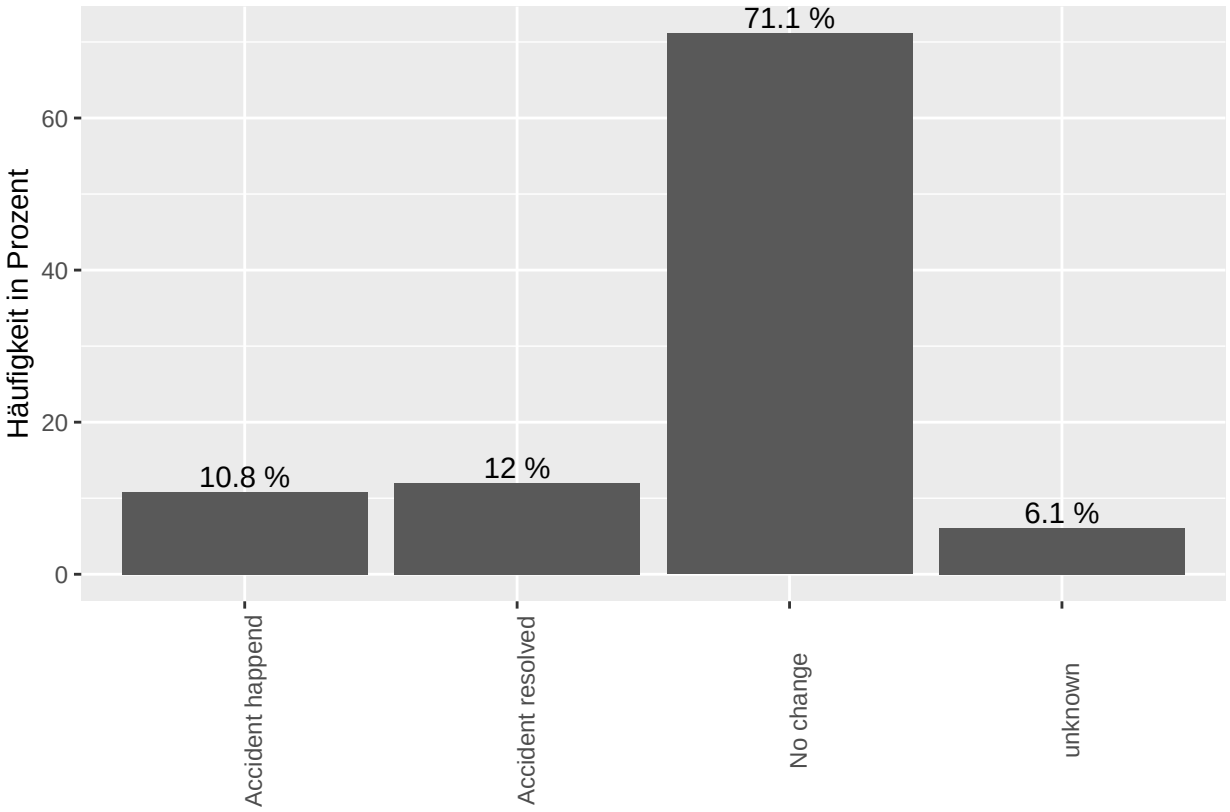

53.5 %
46.5 %
40
20
0
Häufigkeit in Prozent
No
Yes

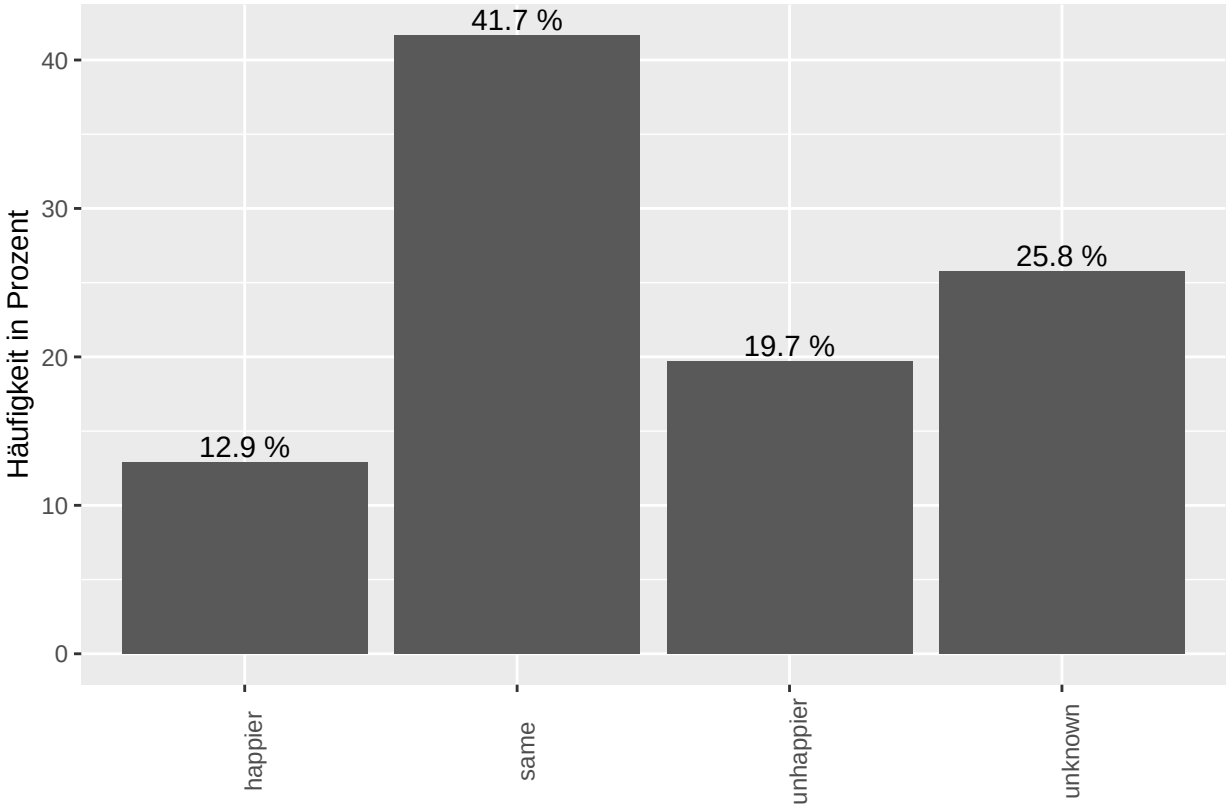

Häufigkeit in Proze
30
20
10
0
3.5 %
8.9 %
0.2 %
28.2 %
0.2 %
0.2 %
18.3 %
4.5 %
0.2 %
0.3 %
0.4 %
0.9 %
Consensual couple with children
Consensual couple without children
Consensual couple without children and another person
Married couple with children
Married couple with children and another person
Married couple without children
Married couple without children and another person
One−person households
One parent with children
One parent with children and another person
Other types of households with and without related family
Other types of households with only related family
Other types of households without related family

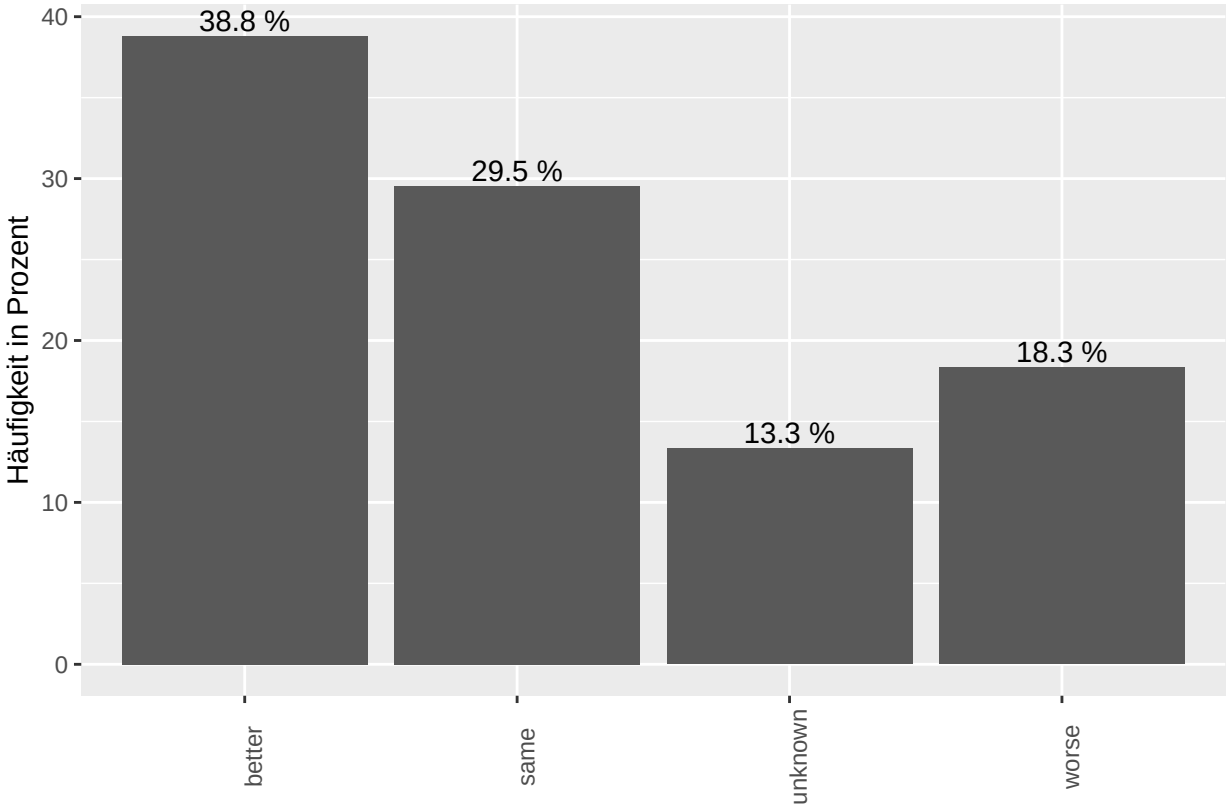

Häufigkeit in Prozent
30
20
10
0
31.6 %
31.5 %
17.9 %
7.7 %
5.3 %
3.8 %
1.4 %
0.8 %
0.1 %
<10000
>10000
>20000
>40000
>60000
>80000
>100000
>150000
>2000000


79.1 %
20.9 %
Häufigkeit in Prozent
80
60
40
20
0
No
Yes


100
75
50
25
0
Häufigkeit in Prozent
95.5 %
4.5 %
No
Yes

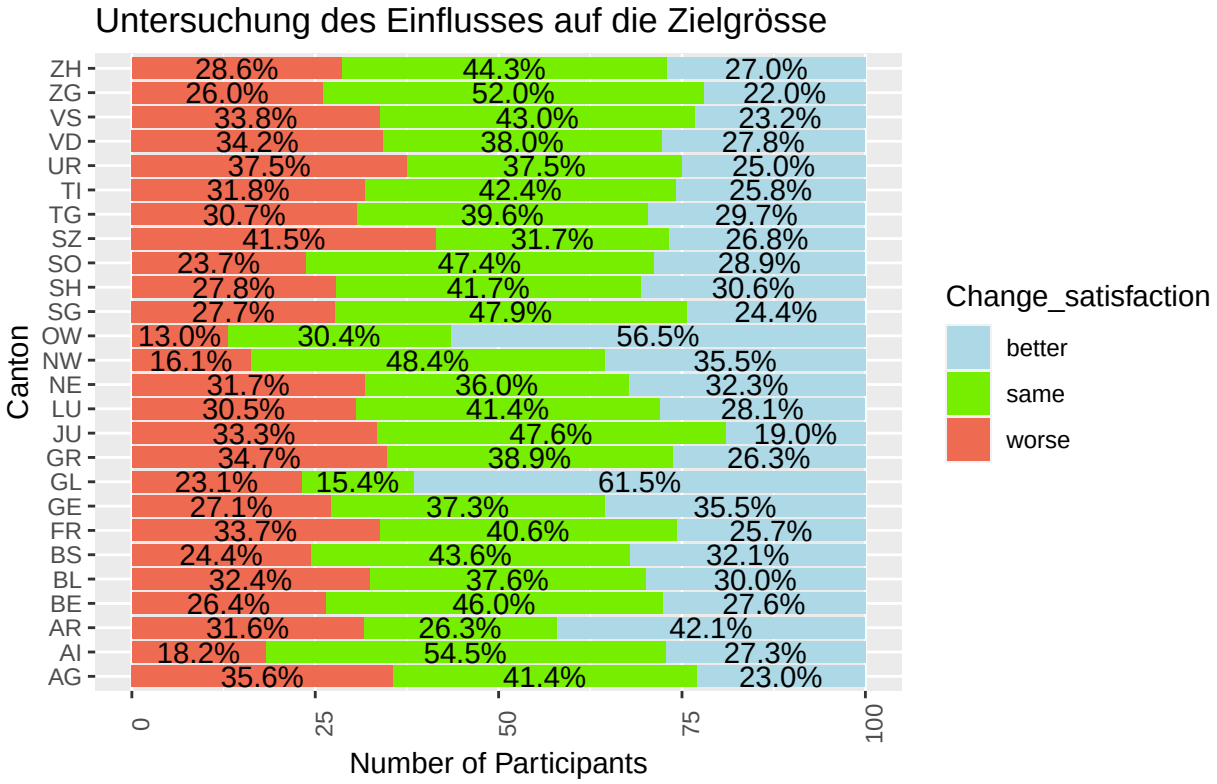

Häufigkeit in Prozent
0
20
40
60
10.8 %
12 %
71.1 %
6.1 %
Accident happend
Accident resolved
No change
unknown


Häufigkeit in Prozent
40
30
20
10
0
41.7 %
25.8 %
19.7 %
12.9 %
happier
same
unhappier
unknown

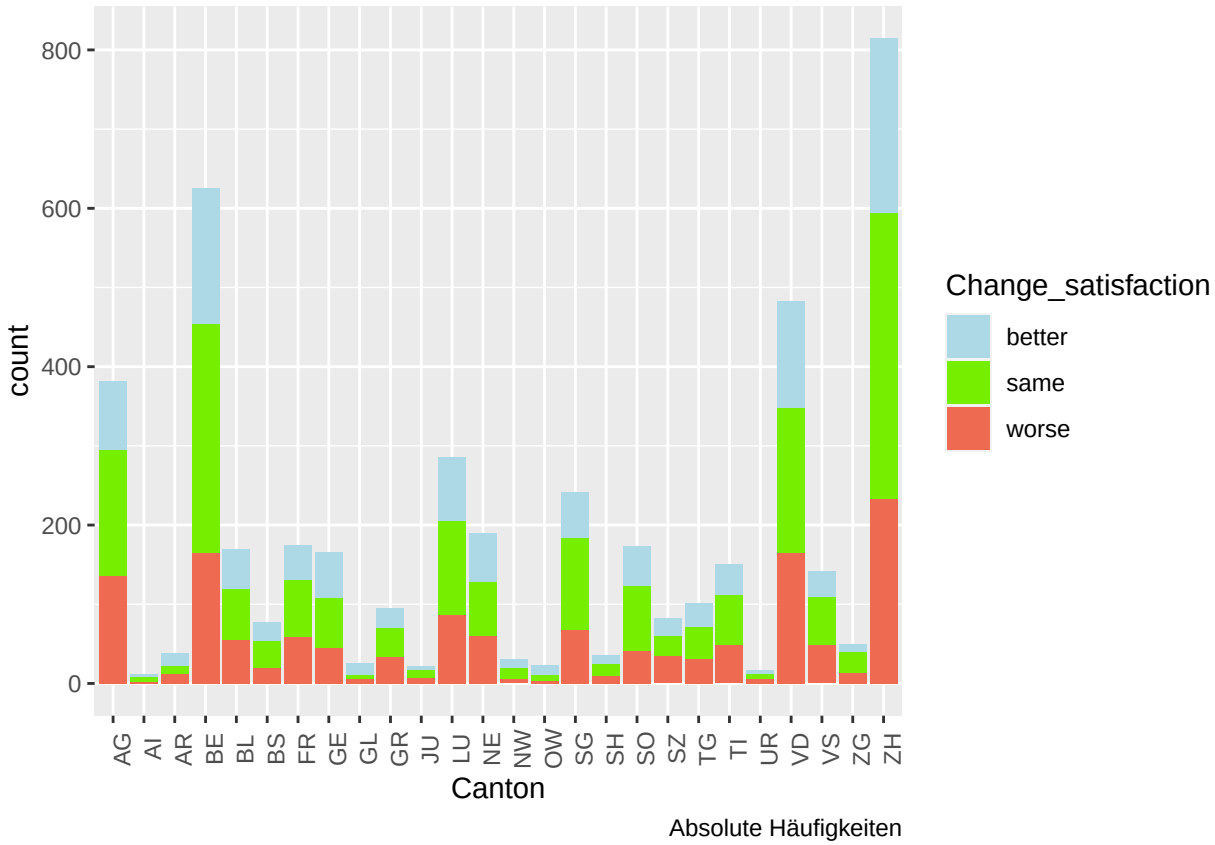

Häufigkeit in Prozent
40
30
20
10
0
38.8 %
29.5 %
13.3 %
18.3 %
better
same
unknown
worse

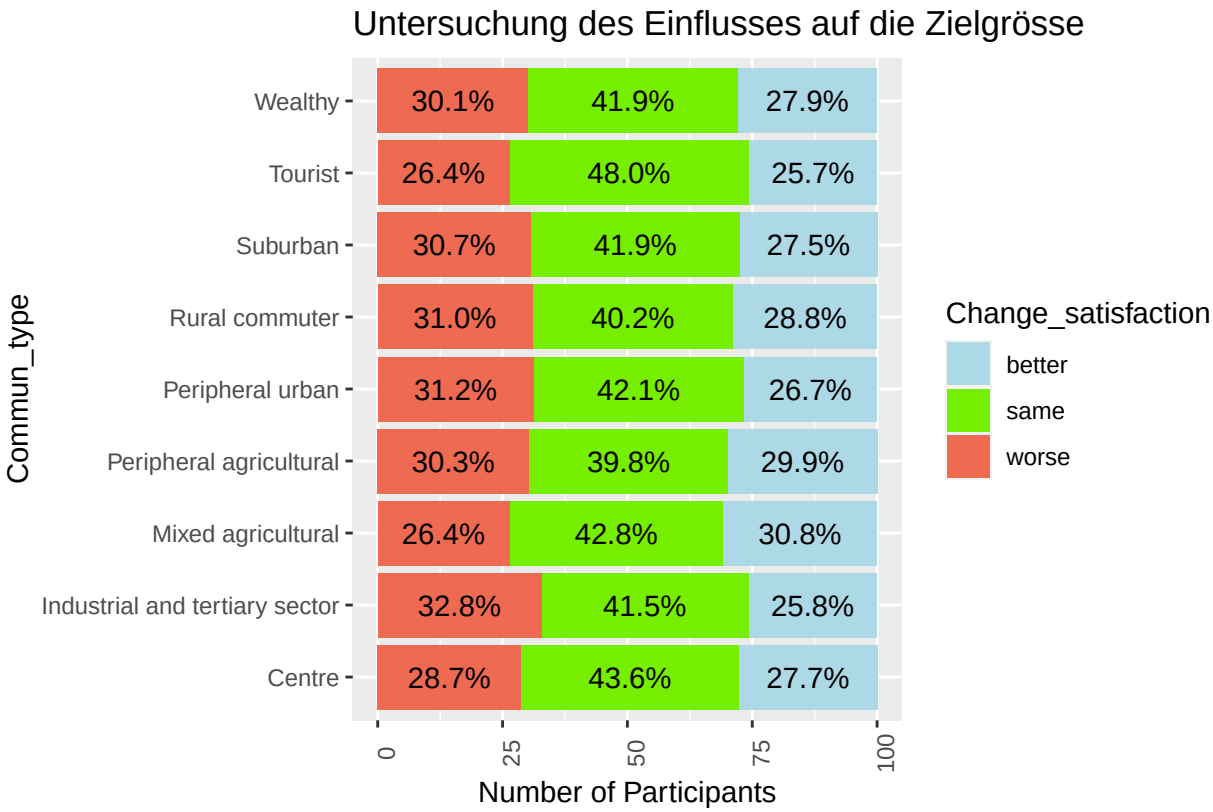
Bivariate Grafiken

Die nachstehenden Abbildungen sollen mögliche Zusammenhänge zwischen den Variablen und der Zielgrösse veranschaulichen.

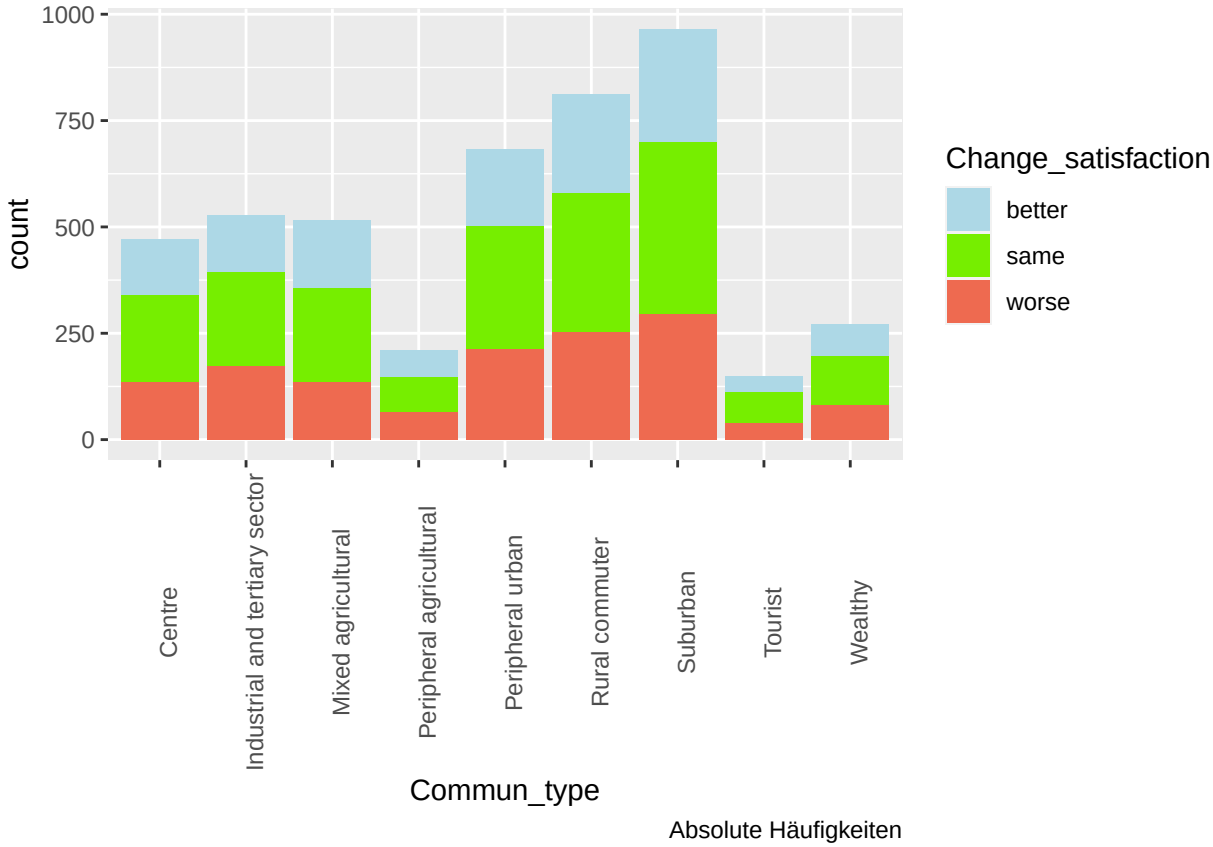


Im Kanton Glarus ist die stärkste positive Veränderung sichtbar, im Kanton Schwyz trat genau das Gegenteil ein.

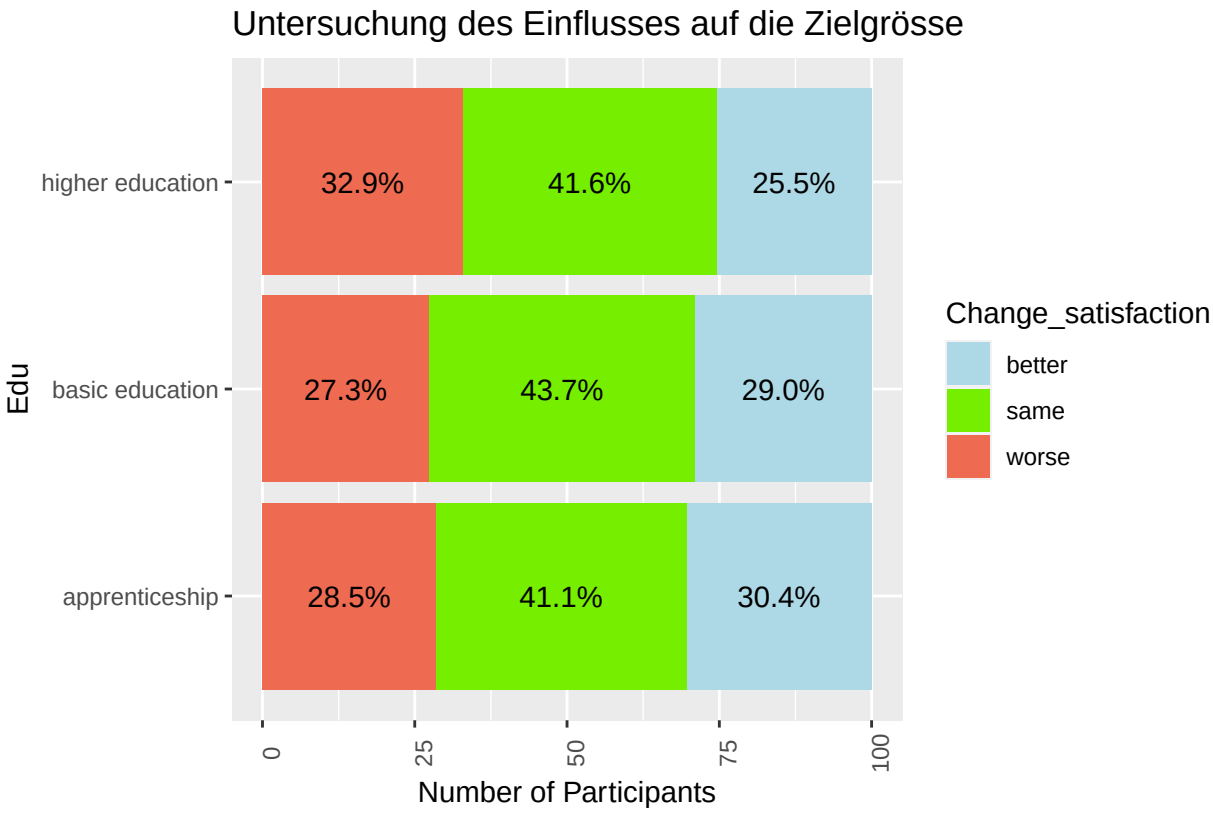

800
600
400
200
0
count
AG
AI
AR
BE
BL
BS
FR
GE
GL
GR
JU
LU
NE
NW
OW
SG
SH
SO
SZ
TG
TI
UR
VD
VS
ZG
ZH
Canton
Change_satisfaction
better
same
worse
Absolute Häufigkeiten



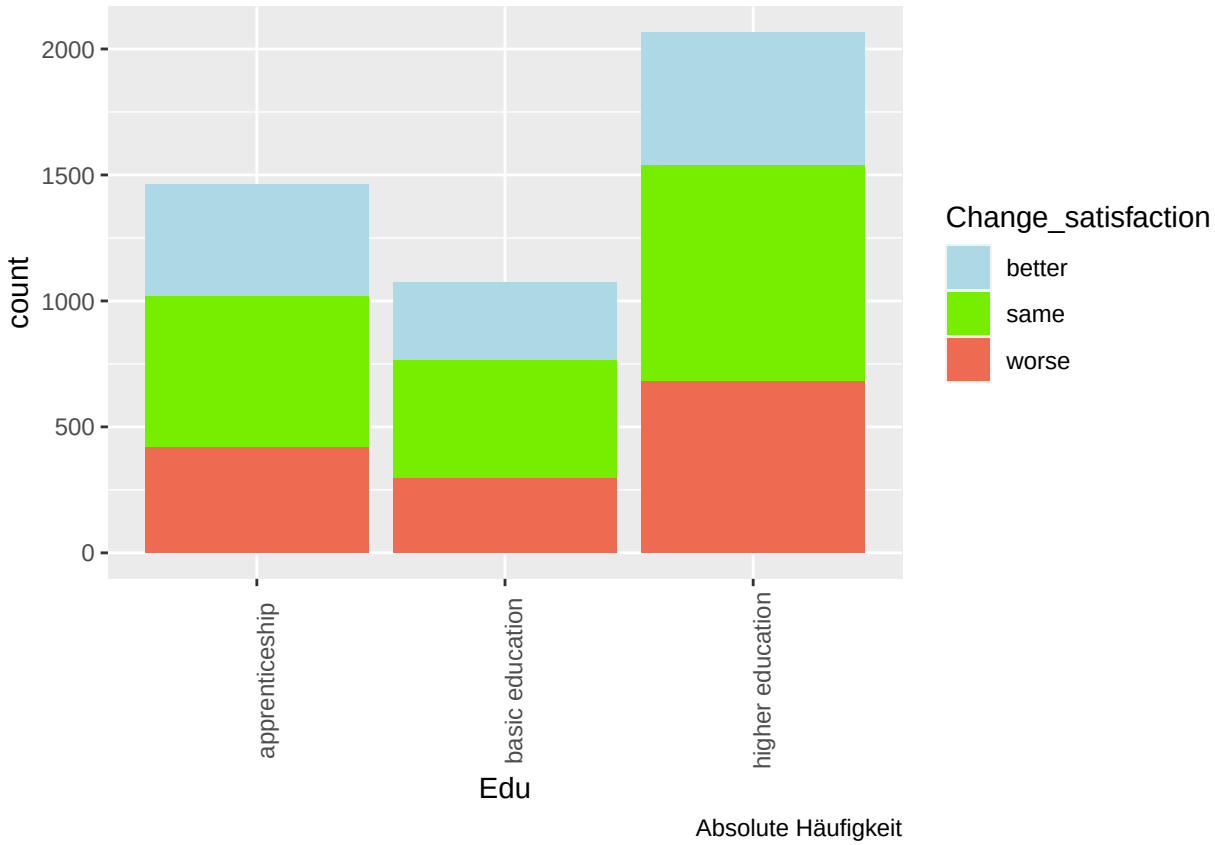
Der Geimeindetyp 'Industrie' hat die grösste negative Veränderung im Bereich der Lebenszufriedenheit.

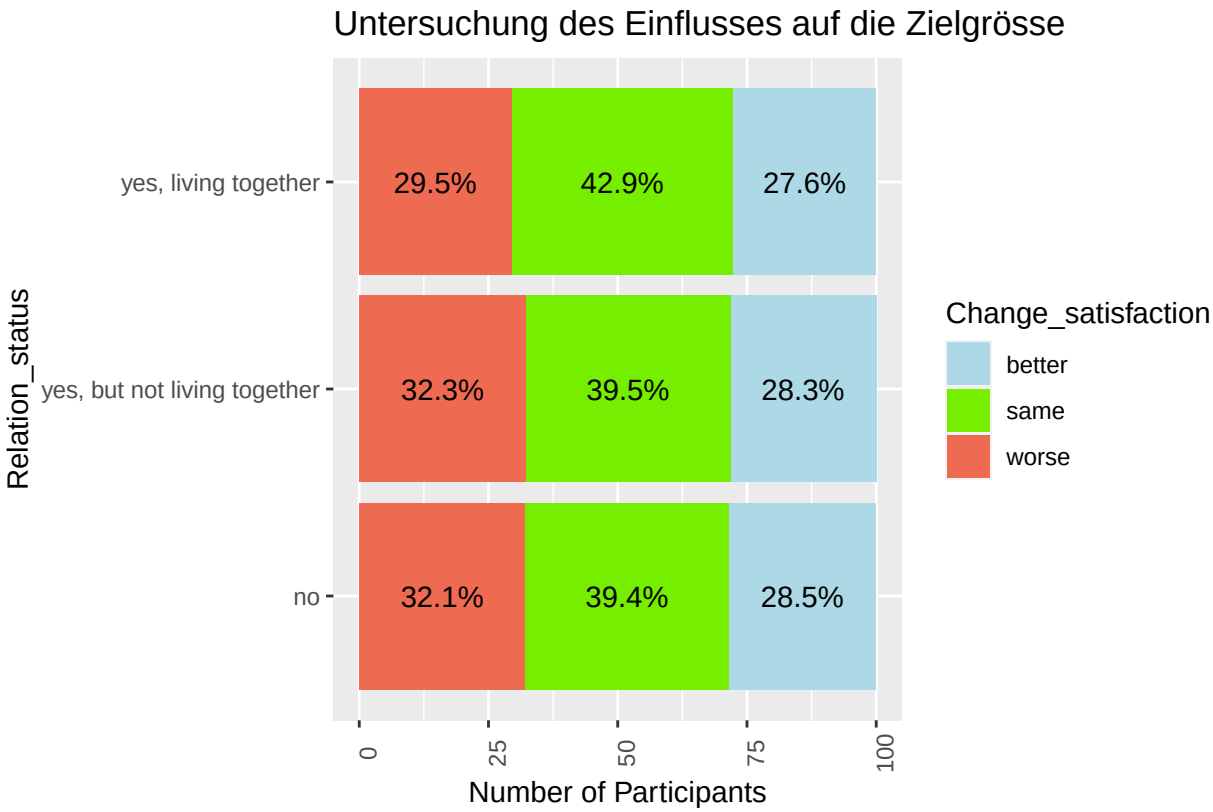


Absolute Häufigkeiten

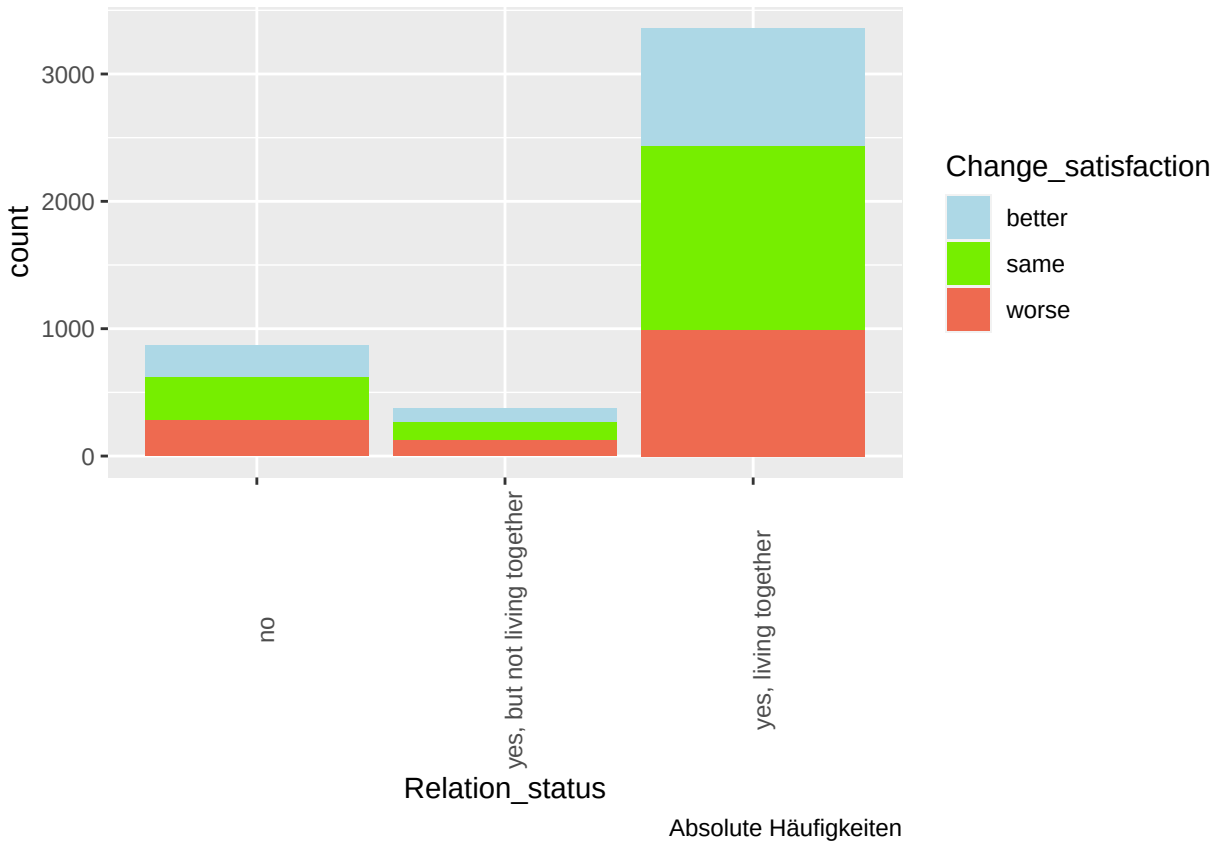


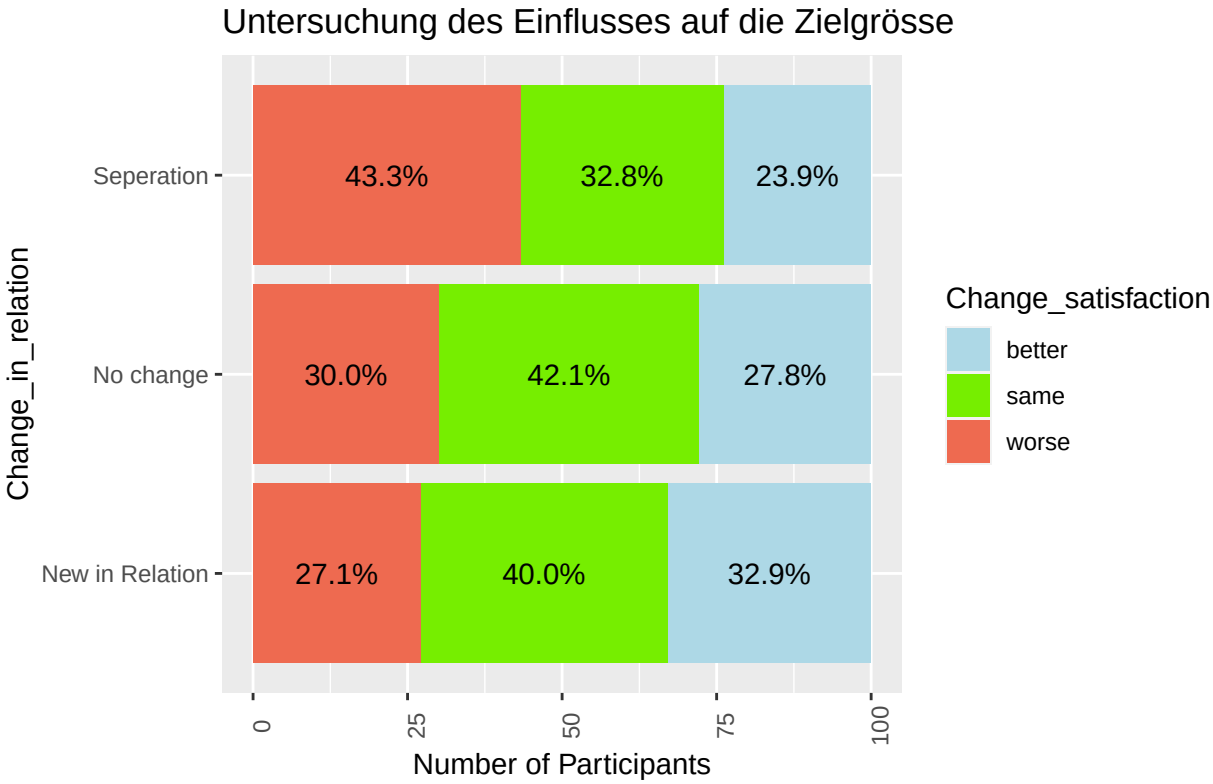
Leute mit tiefen Bildungsstand laut Graphik wurden in der Krise unzufriedener.


2000
1500
1000
500
0
count
apprenticeship
basic education
higher education
Edu
Absolute Häufigkeit
Change_satisfaction
better
same
worse

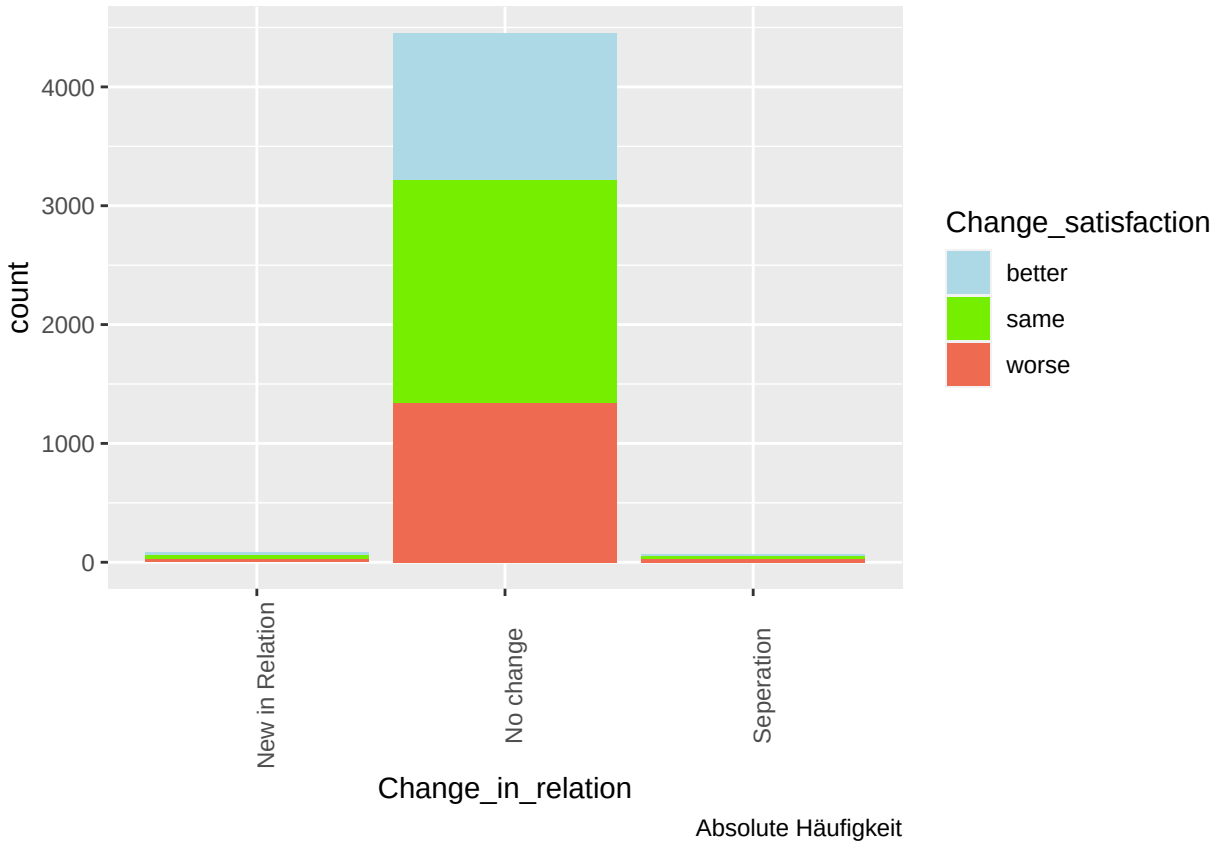


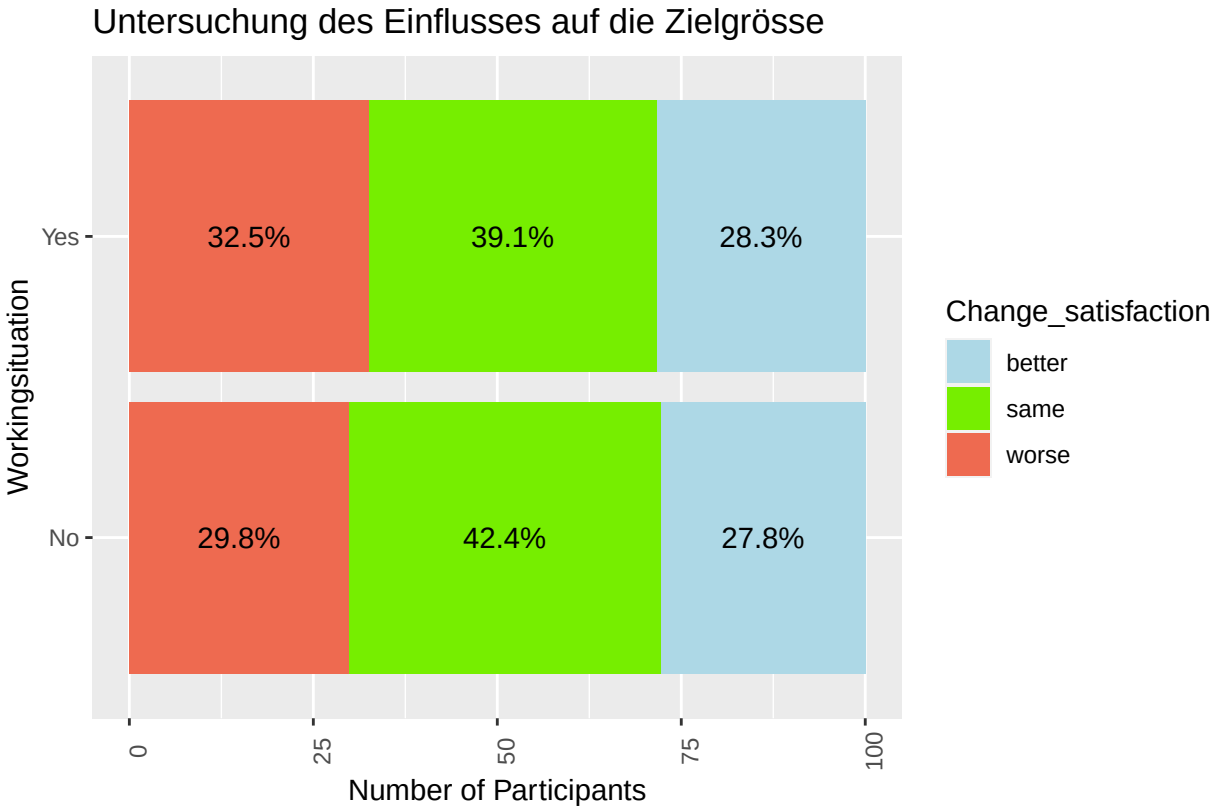
Der Beziehungsstatus hat keinen markanten Einflusses auf die Zufriedenheit


3000
2000
1000
0
count
no
yes, but not living together
yes, living together
Relation_status
Absolute Häufigkeiten
Change_satisfaction
better
same
worse

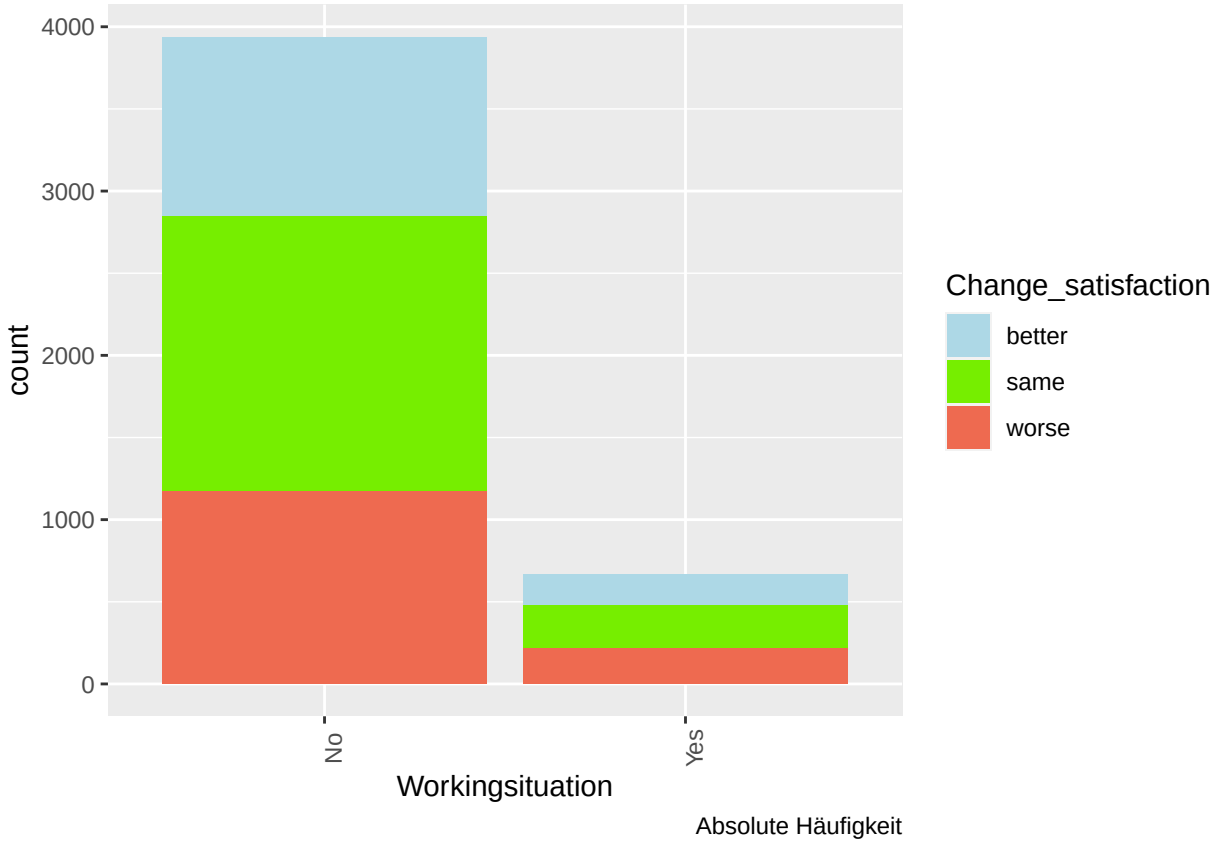


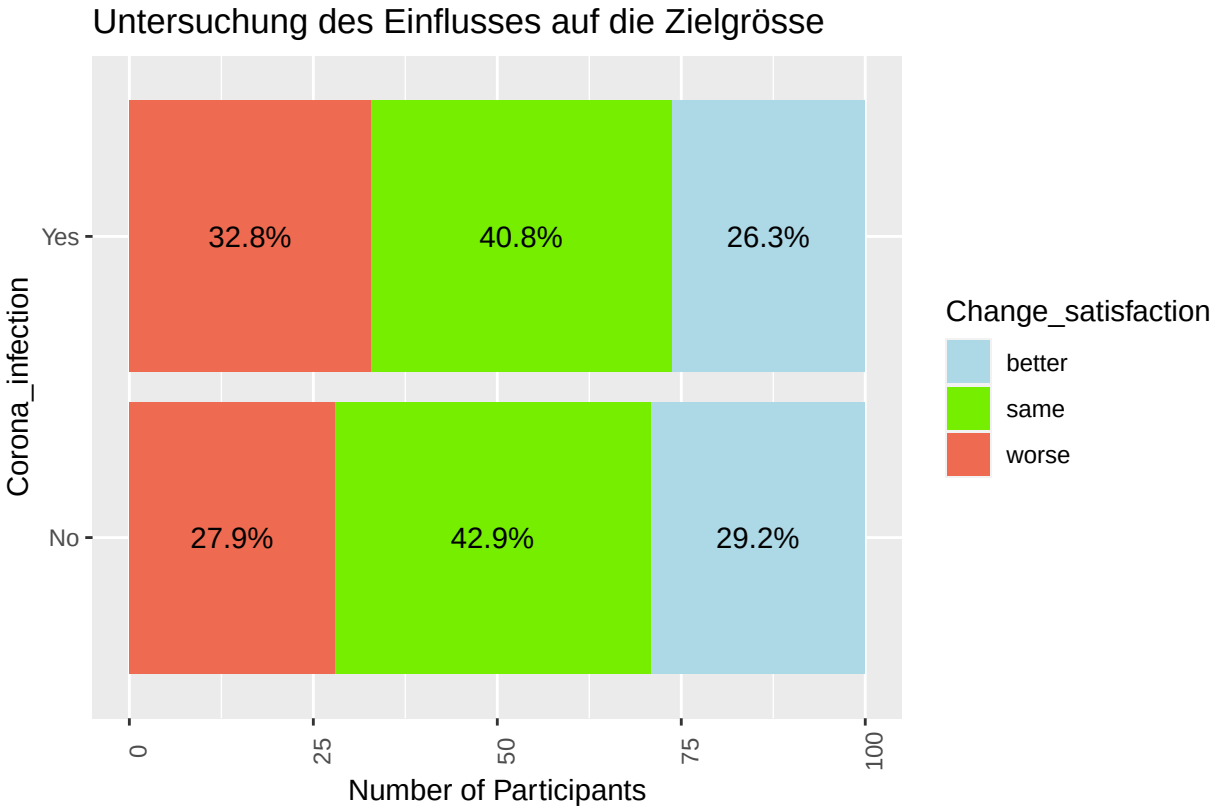
Die Grafik zeigt, dass es den Menschen mit einer neuen Beziehung besser geht als den Menschen, die sich von ihrem Partner getrennt haben.


4000
3000
2000
1000
0
count
New in Relation
No change
Seperation
Change_in_relation
Absolute Häufigkeit
Change_satisfaction
better
same
worse

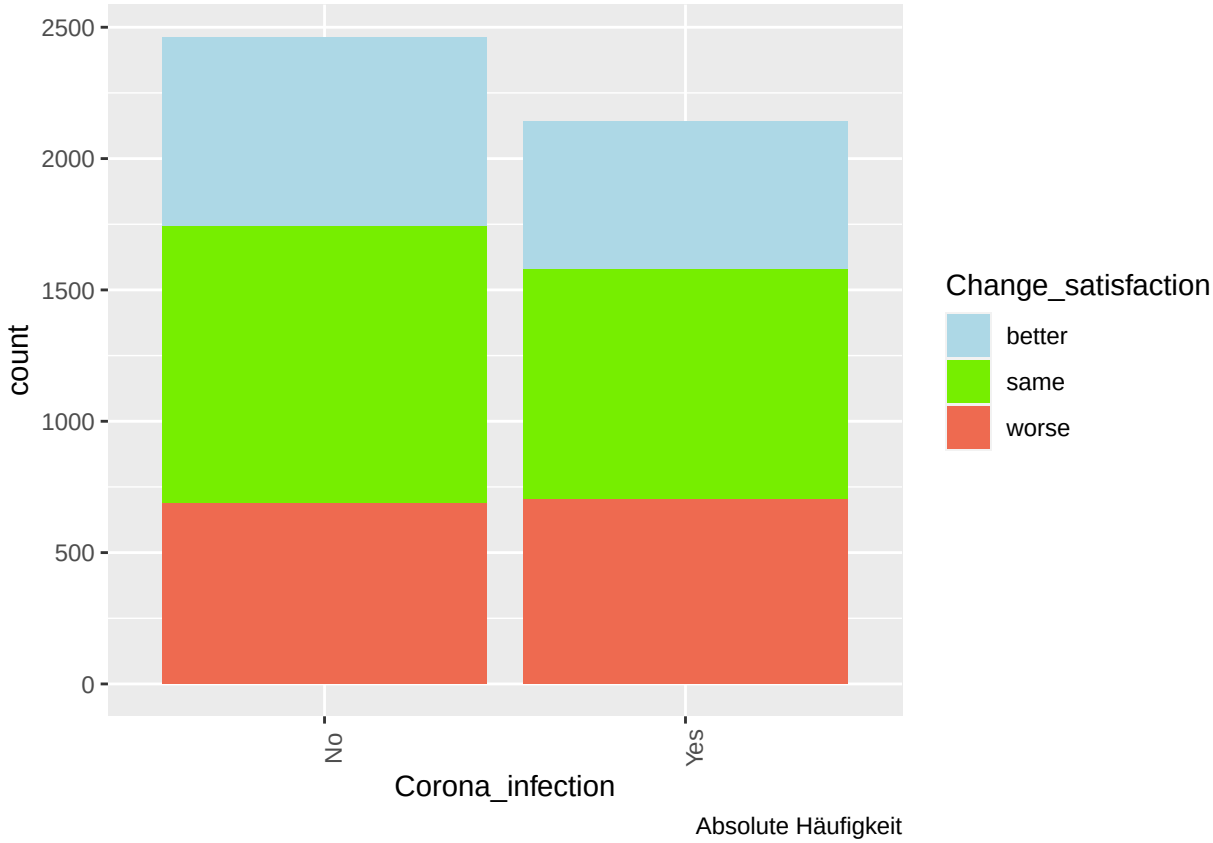


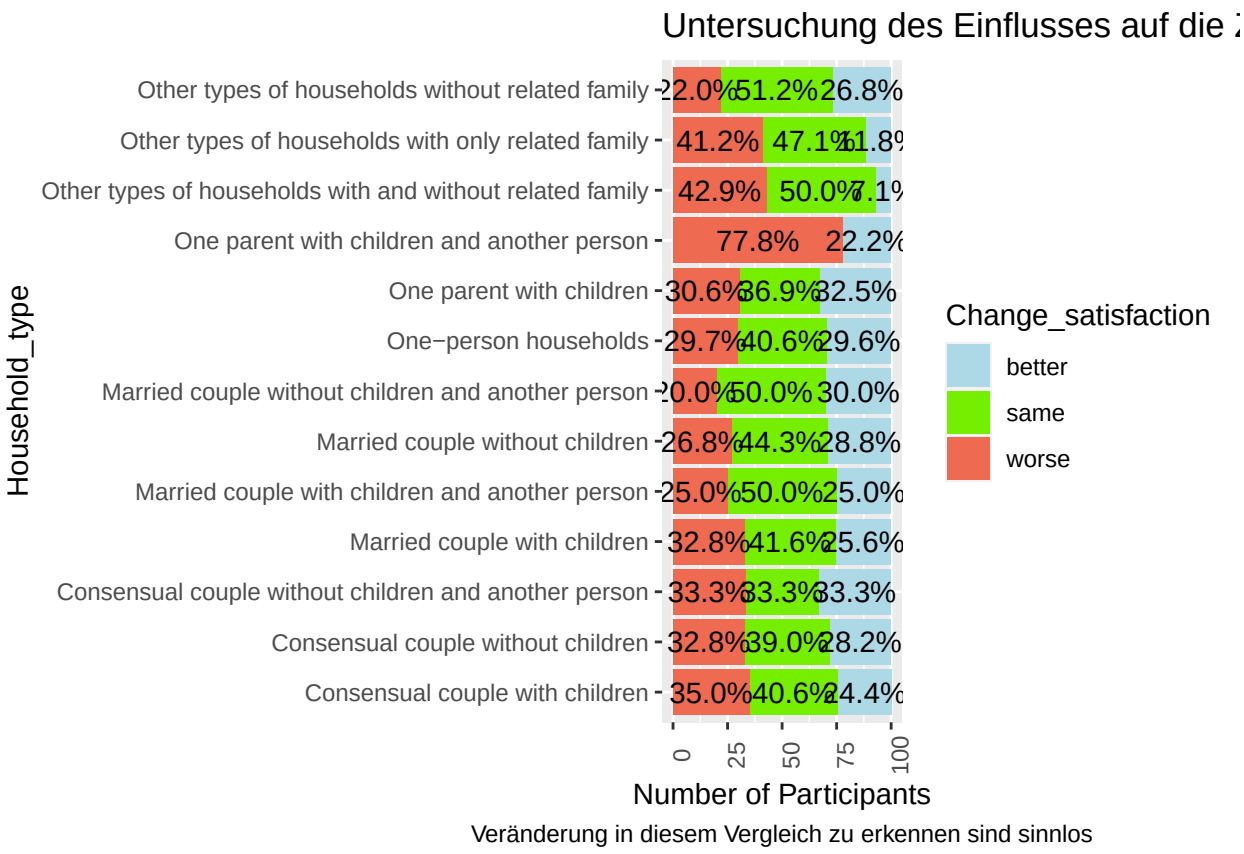
Die Änderung der Arbeitsituation hat keinen (prozentual gesehen) Einfluss

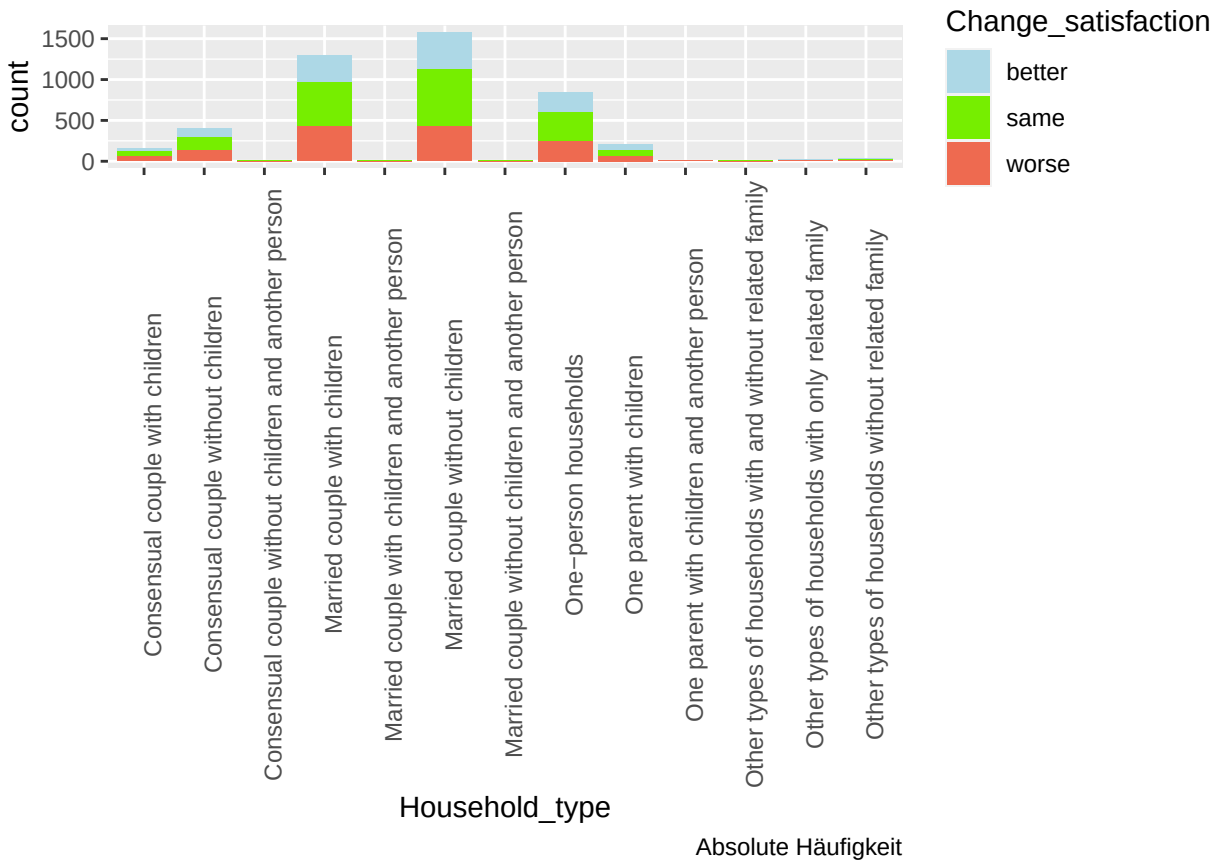

4000
3000
2000
1000
0
count
No
Yes
Workingsituation
Absolute Häufigkeit
Change_satisfaction
better
same
worse

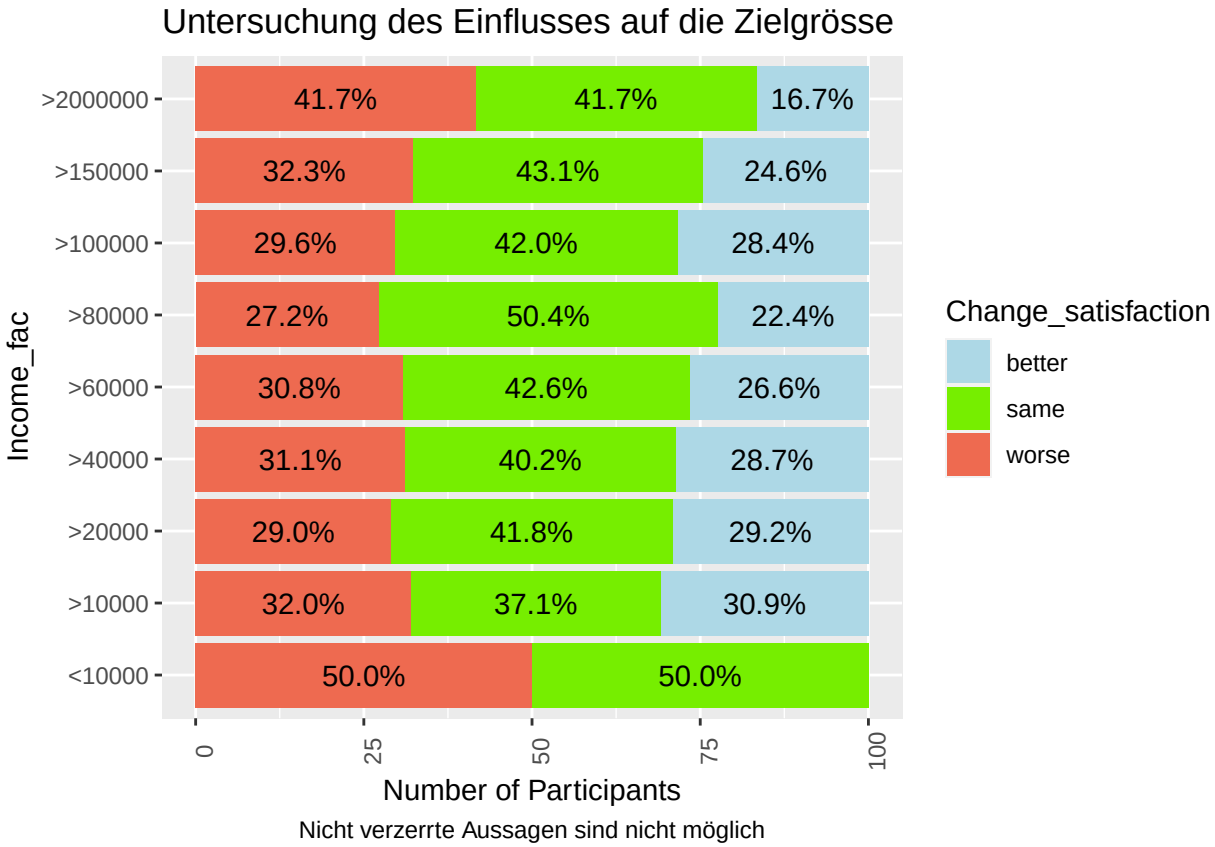


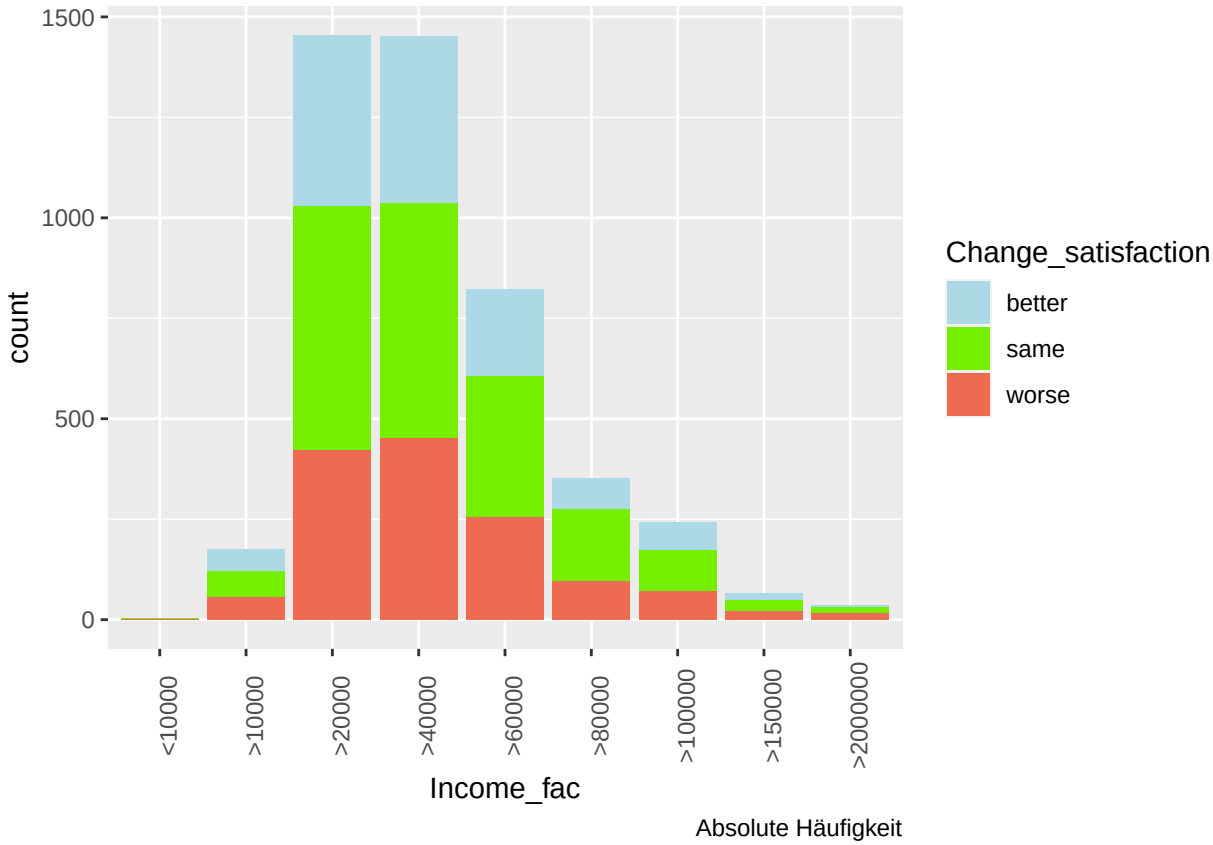
Menschen mit einer Corona−Infektion büssten in Lebenszufriedenheit ein

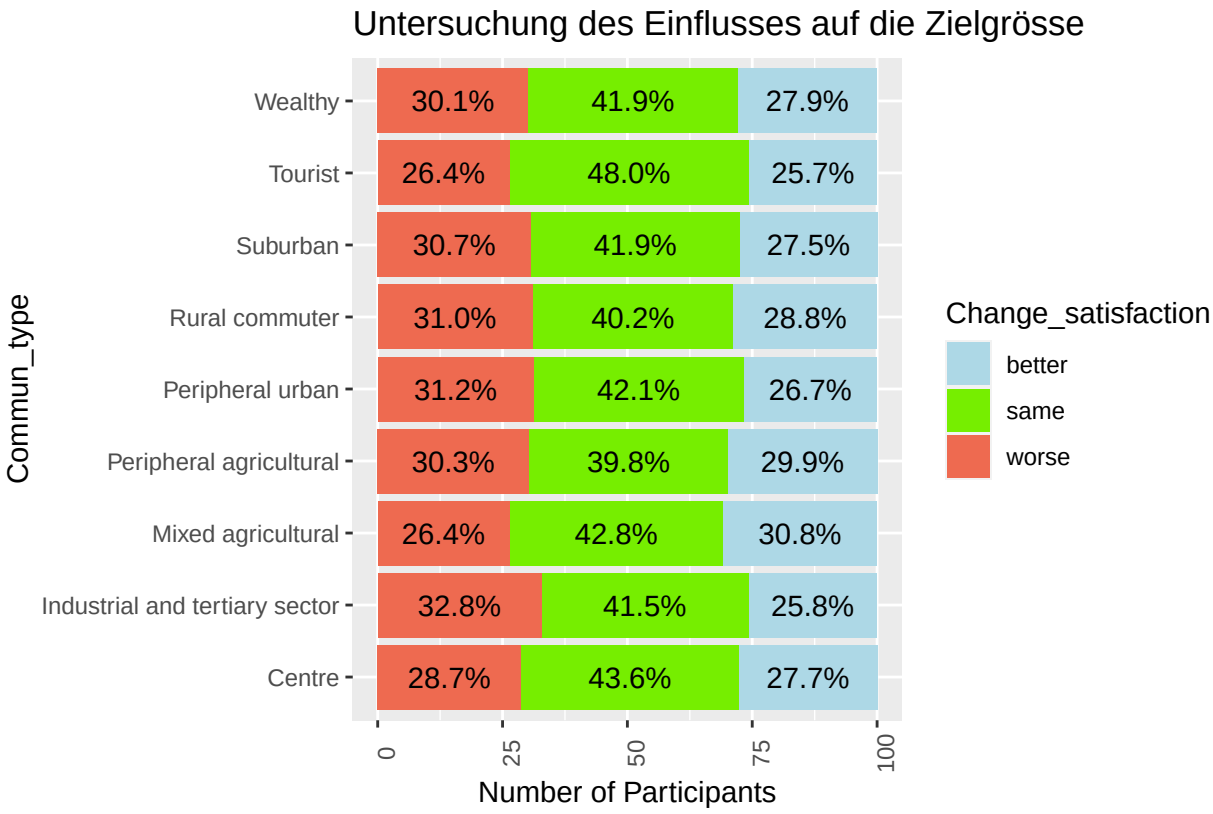

2500
2000
1500
1000
500
0
count
No
Yes
Corona_infection
Absolute Häufigkeit
Change_satisfaction
better
same
worse


Untersuchung des Einflusses auf die
Household_type
Other types of households without related family
Other types of households with only related family
Other types of households with and without related family
One parent with children and another person
One parent with children
One−person households
Married couple without children and another person
Married couple without children
Married couple with children and another person
Married couple with children
Consensual couple without children and another person
Consensual couple without children
Consensual couple with children
22.0%
51.2%
26.8%
41.2%
47.1%
11.8%
42.9%
50.0%
7.1%
77.8%
22.2%
30.6%
36.9%
32.5%
29.7%
40.6%
29.6%
20.0%
50.0%
30.0%
26.8%
44.3%
28.8%
25.0%
50.0%
25.0%
32.8%
41.6%
25.6%
33.3%
33.3%
33.3%
32.8%
39.0%
28.2%
35.0%
40.6%
24.4%
0
25
50
75
100
Number of Participants
Change_satisfaction
better
same
worse
Veränderung in diesem Vergleich zu erkennen sind sinnlos

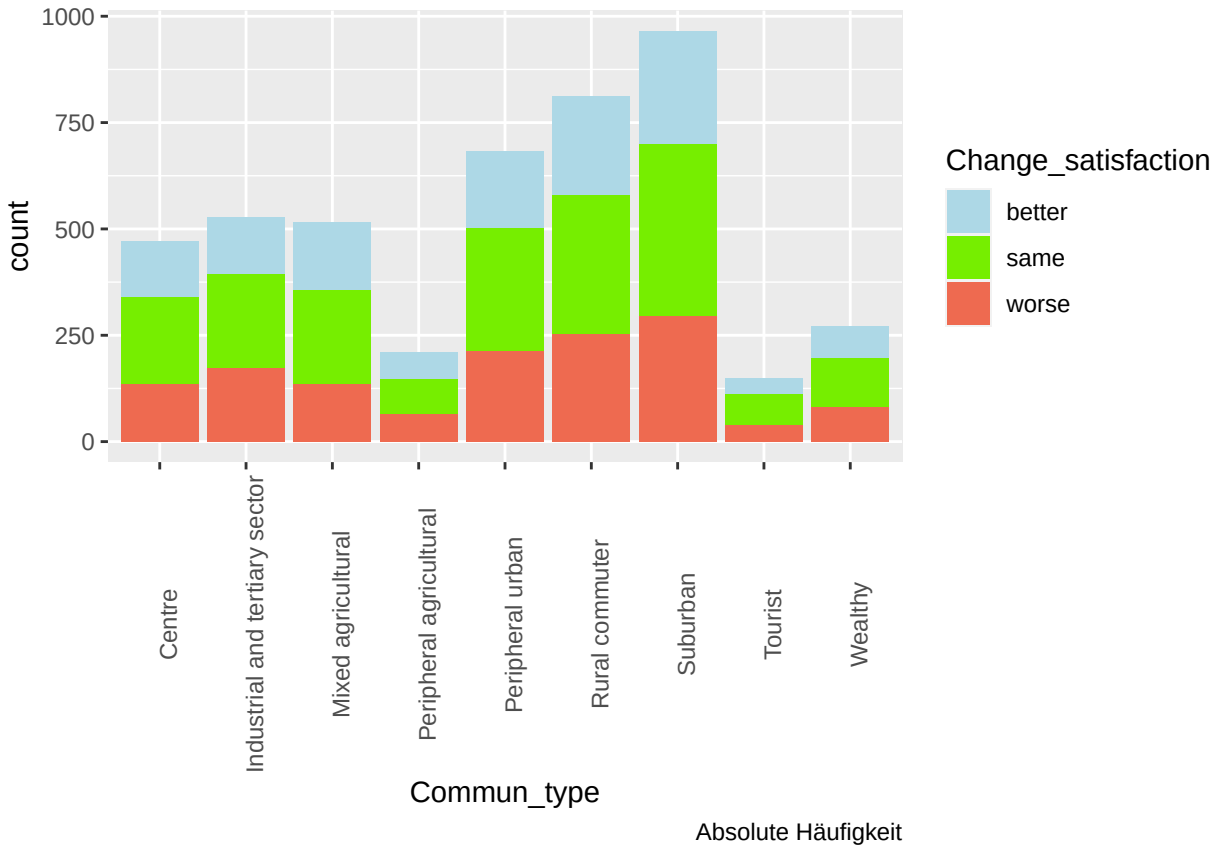

Change_satisfaction
better
same
worse
count
1500
1000
500
0
Consensual couple with children
Consensual couple without children
Consensual couple without children and another person
Married couple with children
Married couple with children and another person
Married couple without children
Married couple without children and another person
One−person households
One parent with children
One parent with children and another person
Other types of households with and without related family
Other types of households with only related family
Other types of households without related family
Household_type
Absolute Häufigkeit

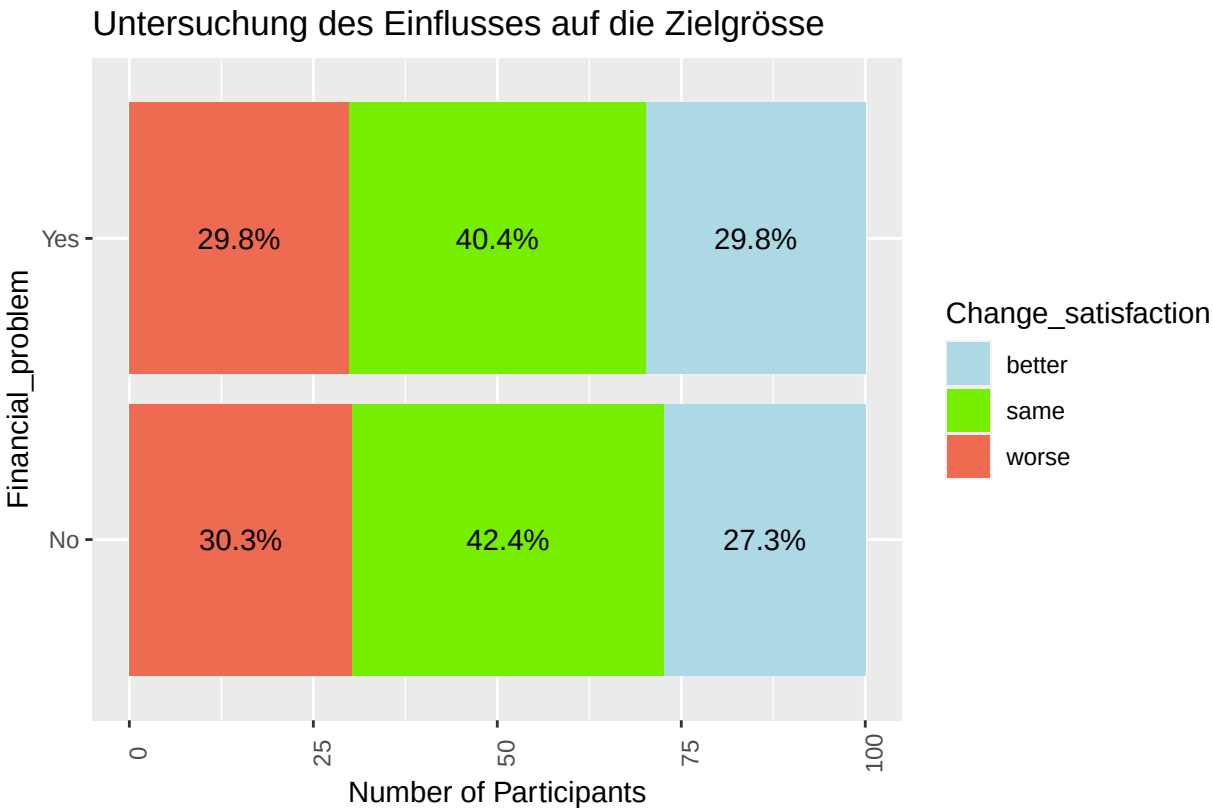

Untersuchung des Einflusses auf die Zielgrösse
Income_fac
>2000000
>150000
>100000
>80000
>60000
>40000
>20000
>10000
<10000
41.7%
41.7%
16.7%
32.3%
43.1%
24.6%
29.6%
42.0%
28.4%
27.2%
50.4%
22.4%
30.8%
42.6%
26.6%
31.1%
40.2%
28.7%
29.0%
41.8%
29.2%
32.0%
37.1%
30.9%
50.0%
50.0%
0
25
50
75
100
Number of Participants
Nicht verzerrte Aussagen sind nicht möglich
Change_satisfaction
better
same
worse


1500
1000
500
0
count
<10000
>10000
>20000
>40000
>60000
>80000
>100000
>150000
>200000
Income_fac
Absolute Häufigkeit
Change_satisfaction
better
same
worse

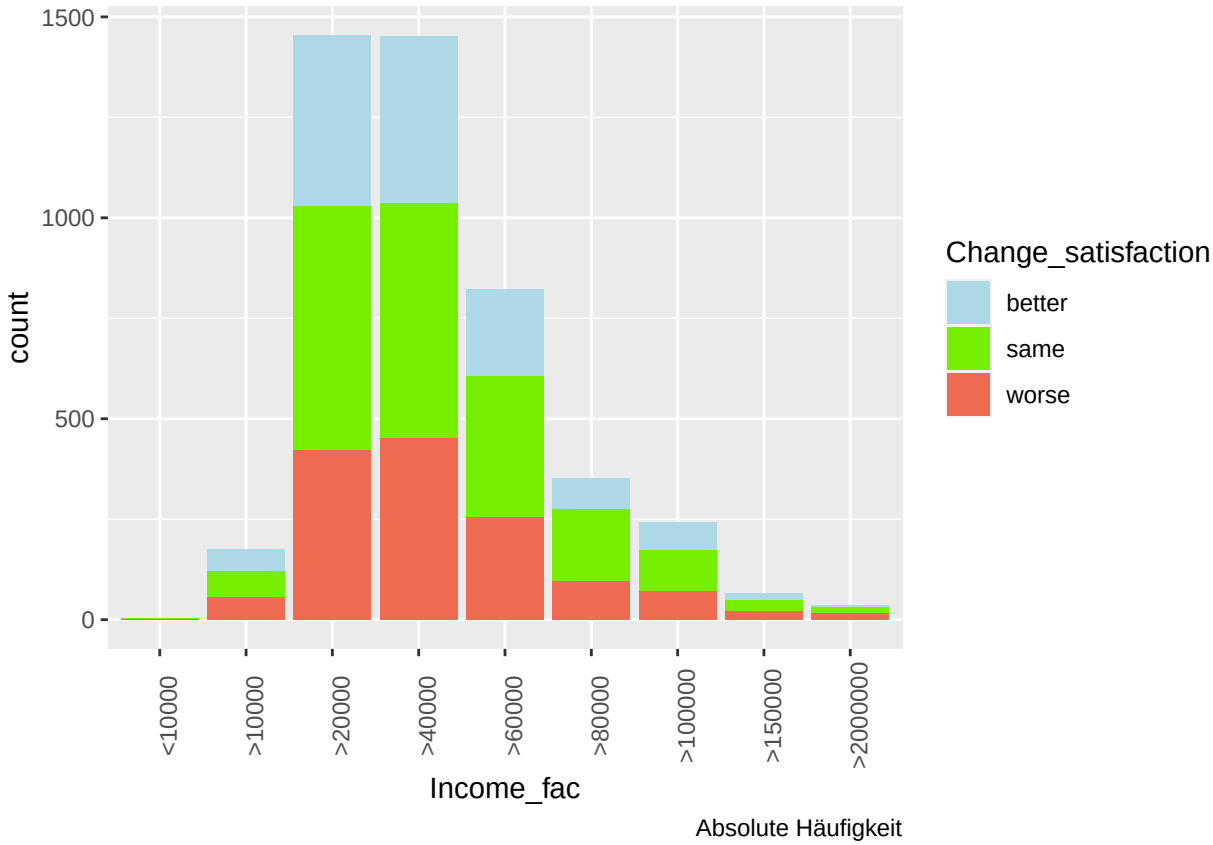


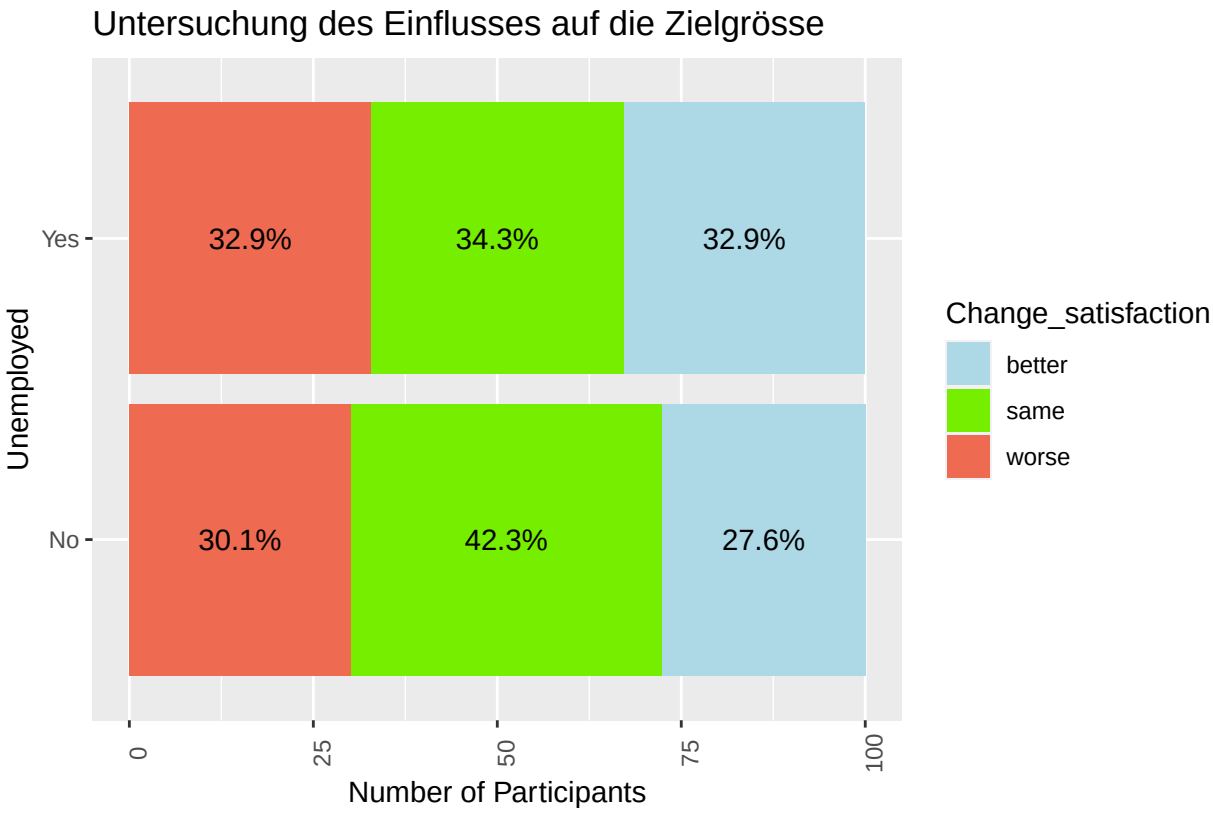
Der Geimeindetyp 'Industrie' hat die grösste negative Veränderung im Bereich der Lebenszufriedenheit


1000
750
500
250
0
count
Centre
Industrial and tertiary sector
Mixed agricultural
Peripheral agricultural
Peripheral urban
Rural commuter
Suburban
Tourist
Wealthy
Commun_type
Change_satisfaction
better
same
worse
Absolute Häufigkeit

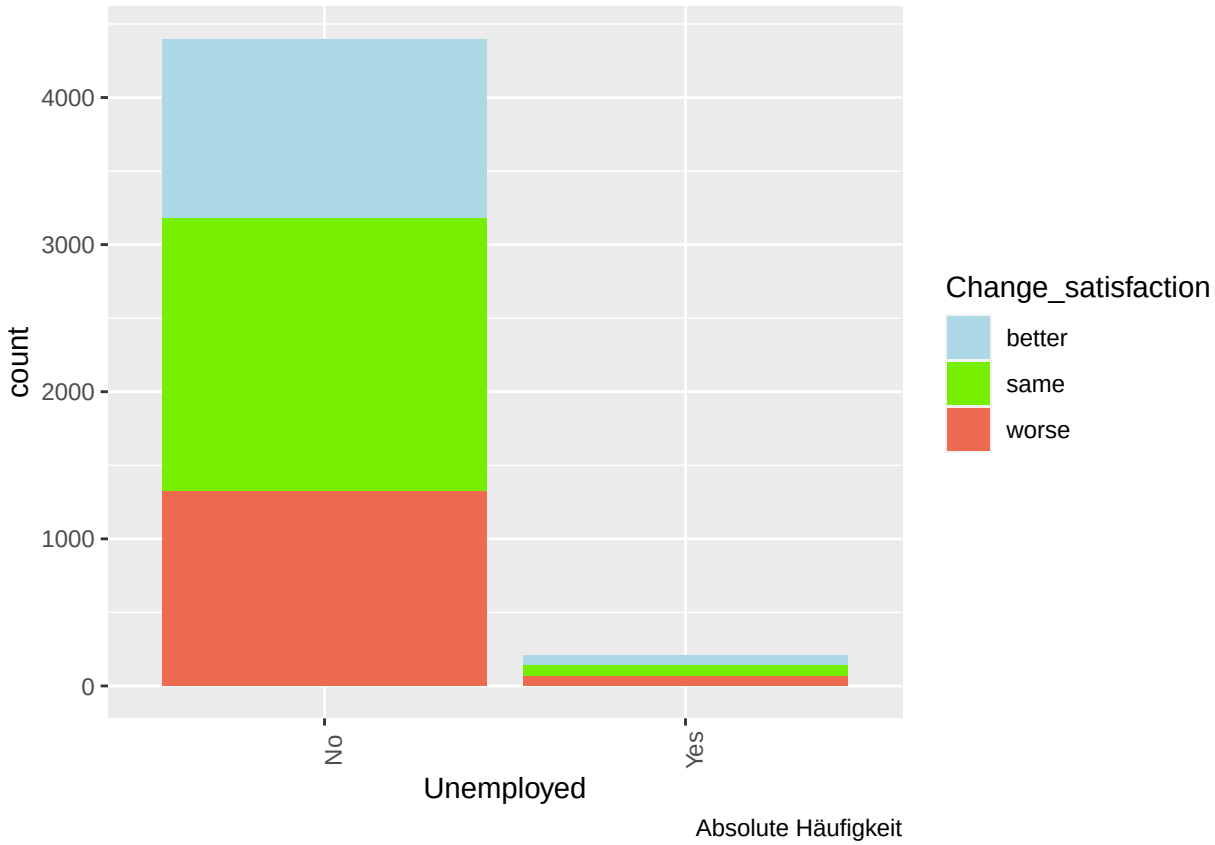


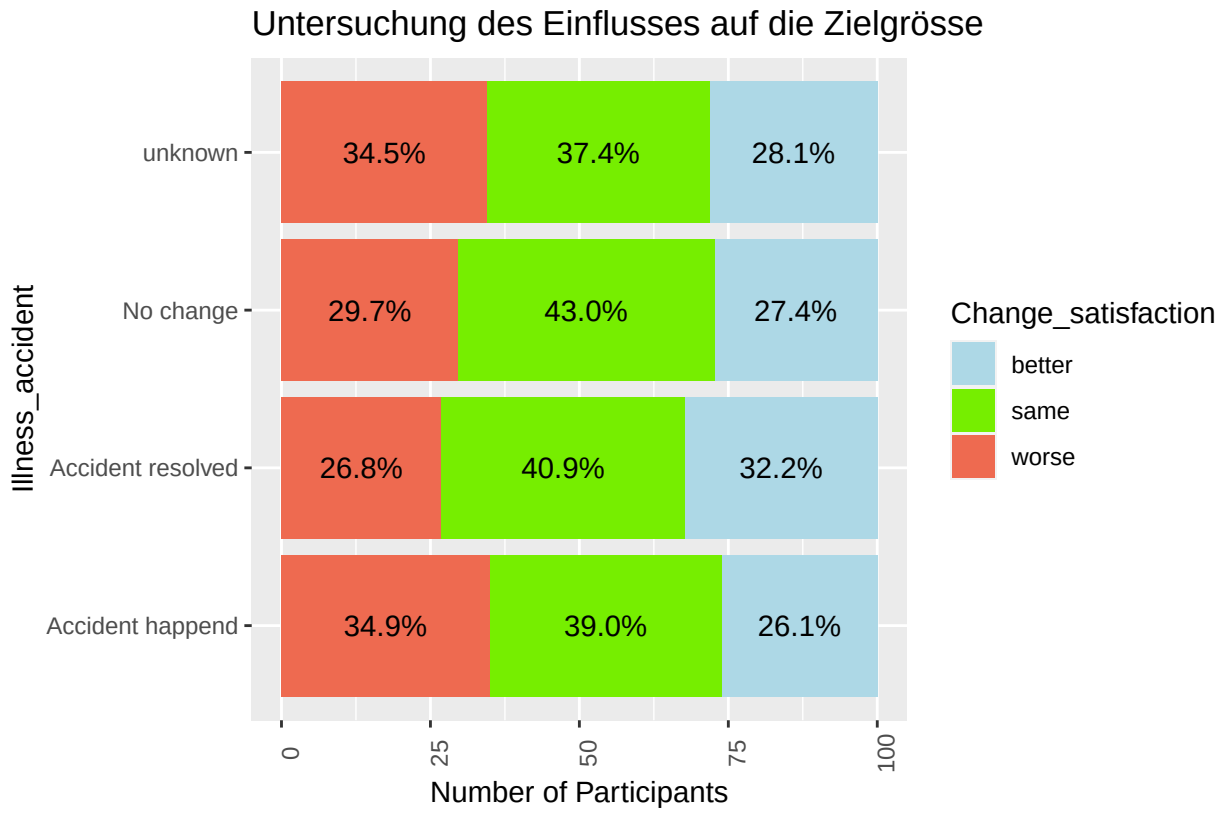
:rstaunlicherweise erging es laut Graphik Leuten mit ehemaligen Finanziellen Problemen besser.


1500
1000
500
0
count
<10000
>10000
>20000
>40000
>60000
>80000
>100000
>150000
>200000
Income_fac
Absolute Häufigkeit
Change_satisfaction
better
same
worse

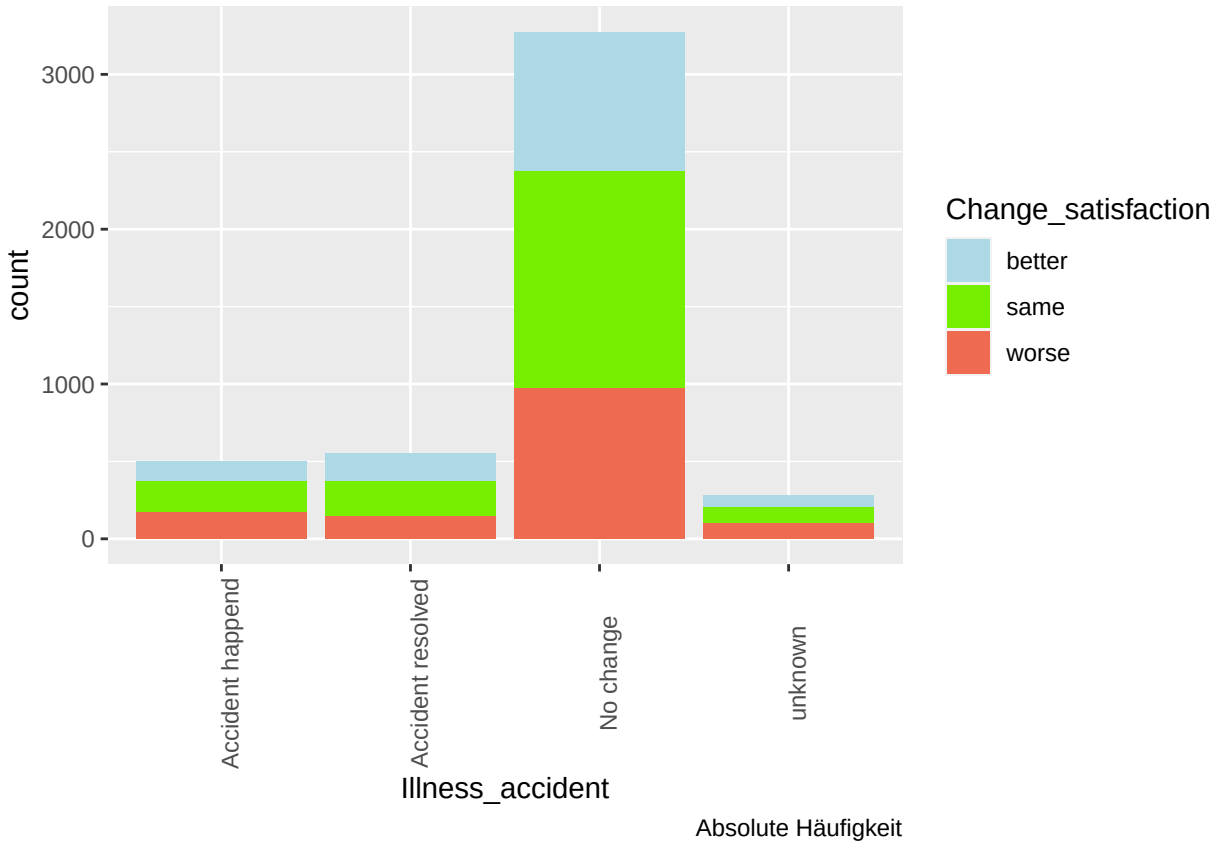


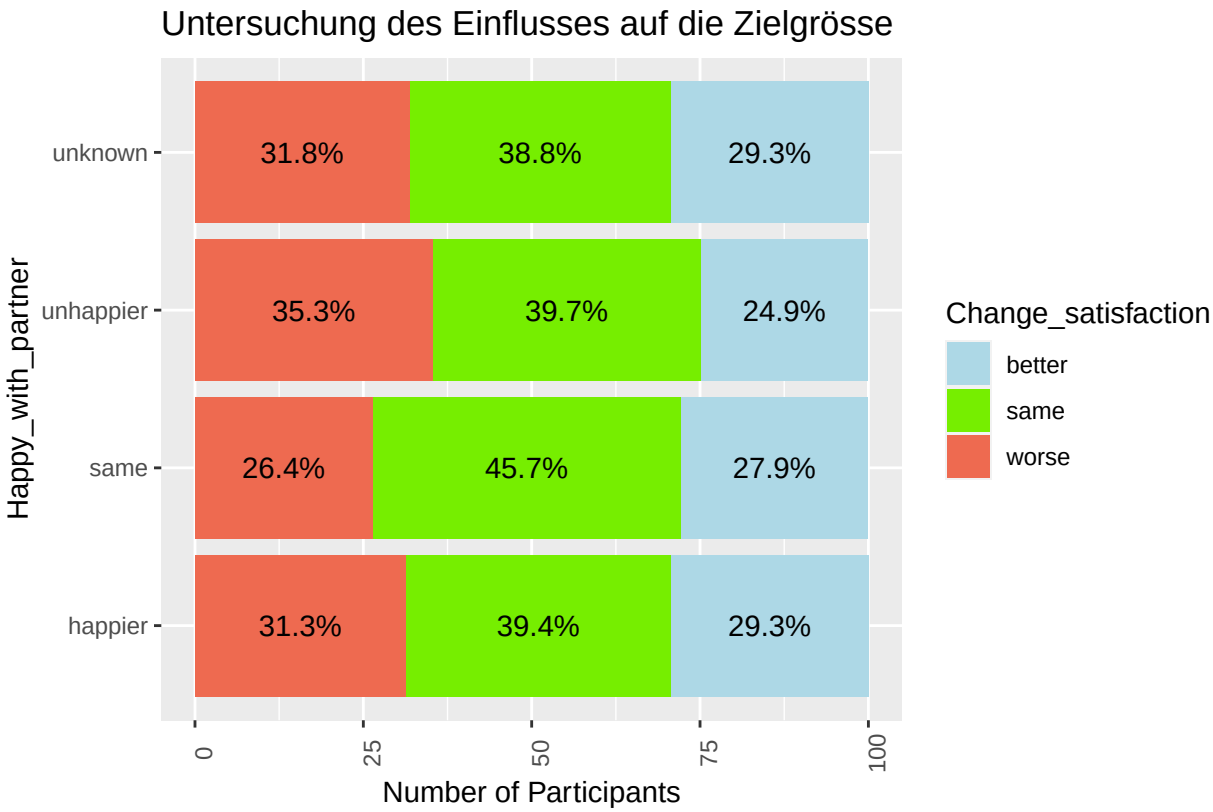
Laut Graphik hat die Arbeitslosgkeit keinen merklichen Einflusses auf die Zielgrösse


4000
3000
2000
1000
0
count
No
Yes
Unemployed
Absolute Häufigkeit
Change_satisfaction
better
same
worse

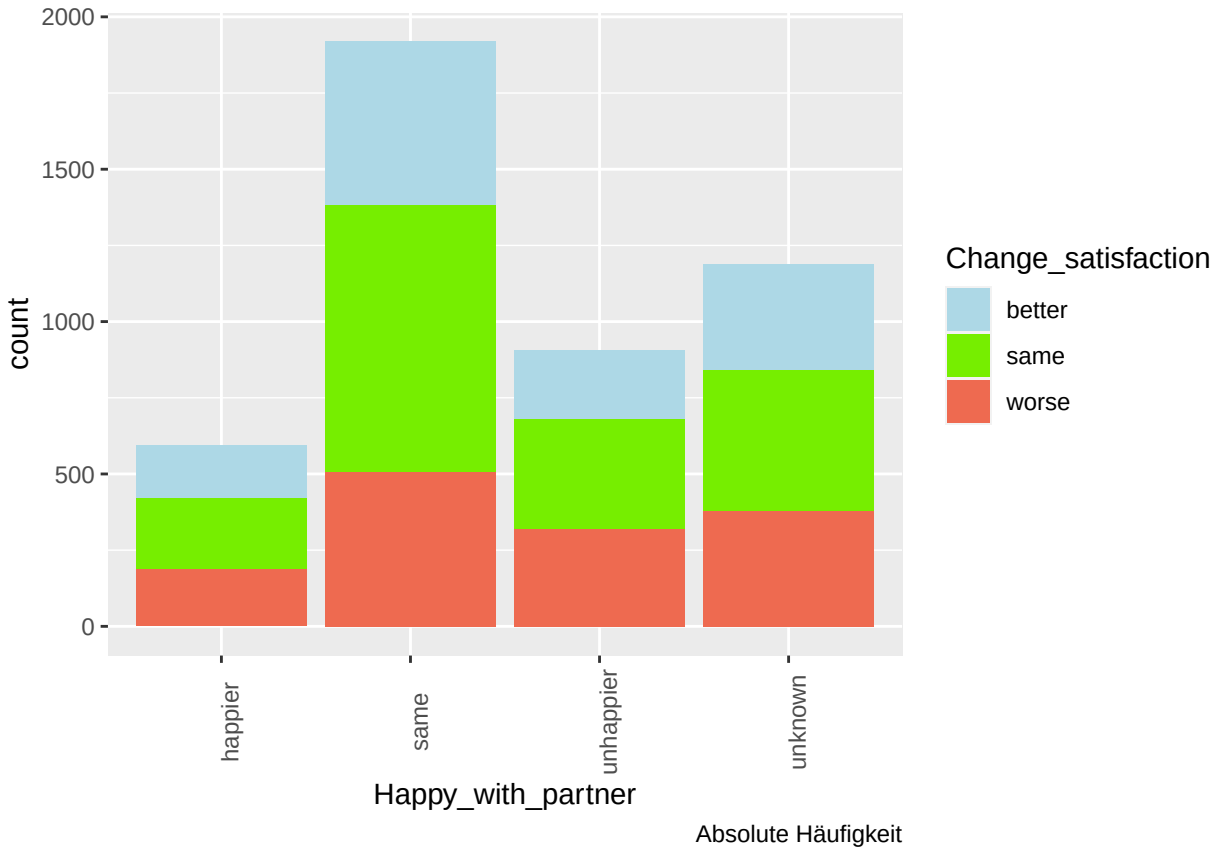


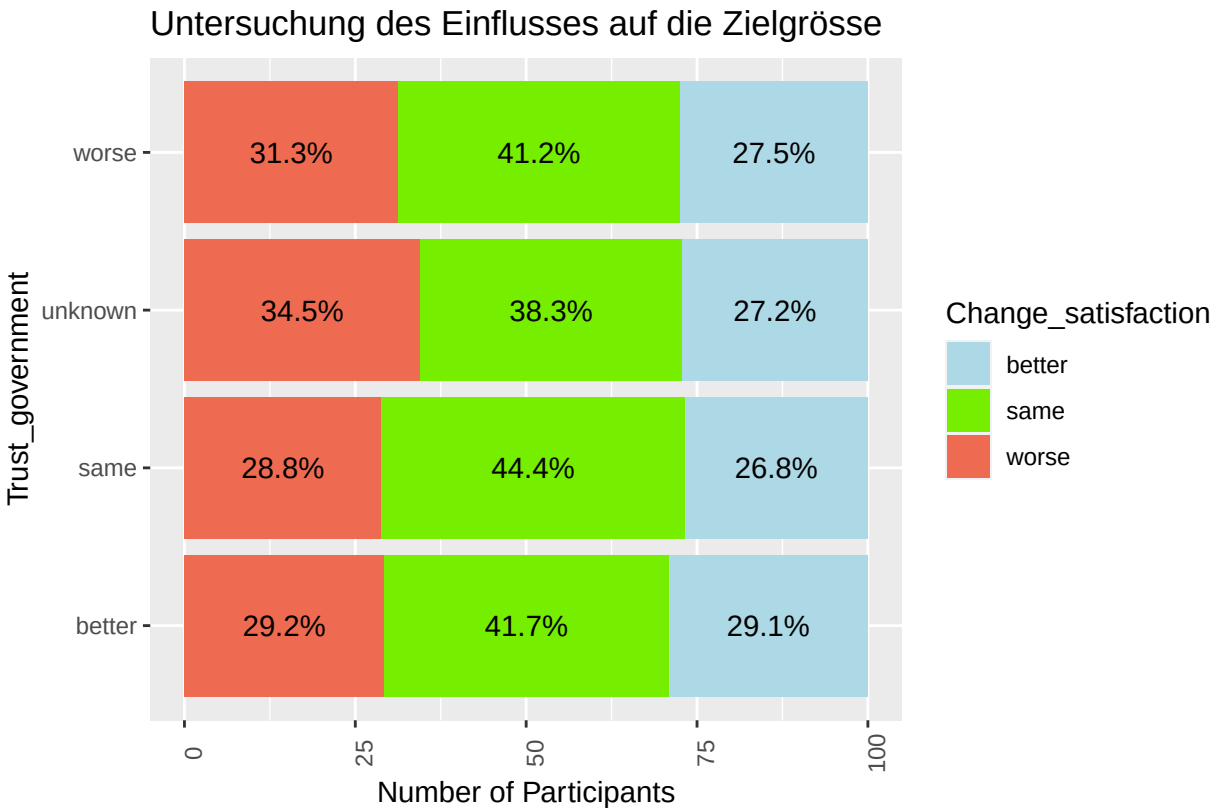
ɪschen, die einen Unfall oder Krankheit erlitten, erfuhren laut Graphik negative Auswirkungen in der Lebenqualität


3000
2000
1000
0
count
Accident happend
Accident resolved
No change
unknown
Illness_accident
Absolute Häufigkeit
Change_satisfaction
better
same
worse

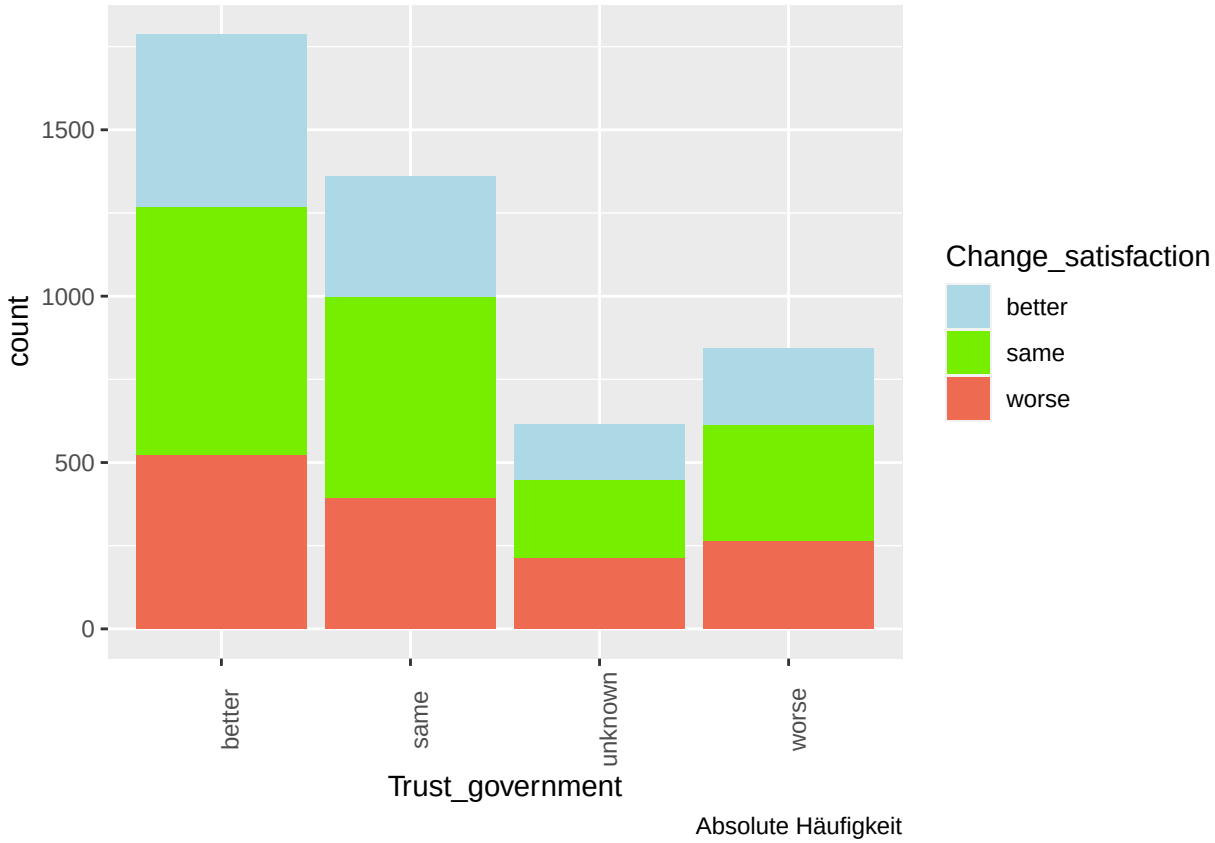


Leute, die unglückich in der Beziehung waren, verloren laut Graphik auch an Lebensqualität.


2000
1500
1000
500
0
count
happier
same
unhappier
unknown
Happy_with_partner
Absolute Häufigkeit
Change_satisfaction
better
same
worse



schen mit Vertrauensverlust in die Regierung erlitten erhöhte negative Auswirkungen in der Lebensqualität



Schlusswort

Die Zielvariable hat sich im Verlauf der Pandemie nicht merklich verändert. Die Lebenszufriedenheit hat sich für etwa $\frac{2}{5}$ der Bevölkerung nicht verändert. Für die restlichen $\frac{3}{5}$ kam es entweder zu einer Verbesserung oder Verschlechterung der Lebenszufriedenheit. Man kann aber nicht den Schluss ziehen, dass es an der Pandemie lag.