Dokumentation Praktikum 2

Bader Eddin Loukil, Nevio Roccia

27 4 2022

Vorwort

Dieses Markdown wurde im Rahmen des Moduls "Data Processing with R" verfasst, und dient zur Dokumentation aller Arbeitsschritte des Praktikum 2 "Swiss Household Data". Hier finden sie zusätzliche Informationen zum gesamten Ablauf der Projektarbeit und aller Arbeitsschritte, so wie eine Diskussion der erzielten Resultate. Ziel dieses Projektes ist es, die Datensätze des Swiss Houshold Panels auszuwerten. Wir verfolgen dabei die Frage: "Welche Faktoren stehen mit einer Veränderung der allgemeinen Lebenszufriedenheit seit Beginn der Verbreitung von Covid-19 in Beziehung?". Eine Übersicht über die Abschnitte der Dokumentation und auch des Codes erhalten sie durch drücken von "Ctrl + Shift + O". (Hier, sowie auch im Code selber) Im Text stehen immer wieder Zeilenangaben in Klammern. Diese weisen auf die respektiven Zeilen im R-Code hin. Bsp: (Zeile 104)

Dokumentation

Planung

Eine grobe Planung und Aufteilung der Arbeitsschritte der Arbeit finden Sie im "README.md"-File. Dort haben wir die Aufgaben jeweils nach den Wochen unterteilt und provisorisch jemandem zugewiesen.

Organisation & Datensicherheit

Unsere Organisation wurde im Abschnitt "Planung"im "README"-File unter "Woche 1 Datenakquise & Lagerung" bereits erwähnt. Da wir gemeinsam am gleichen Datenset arbeiten müssen, haben wir zur Lagerung der Rohdaten ein Gruppe in Microsoft Teams erstellt. Dort sind sie unter "Dateien" in einem Ordner mit dem Titel "Data_SPSS" zu finden. Dies gewährleistet einerseits, unkomplizierten Zugang zu den Daten über eine Software mit der wir bereits vertraut sind, andererseits bietet sie den benötigten Datenschutz, da diese Gruppen für Dritte nicht zugänglich sind. Da das Projekt auf Kollaboration basiert und wir gemeinsam am gleichen Code arbeiten müssen, nutzen wir zusätzlich GitHub um den Code flexibel teilen zu können. Da man die Datensätze jedes Mal neu einlesen muss wenn man am Code arbeitet, entschieden wir uns dazu am Anfang beim einlesen die Daten zweimal einzulesen. Das erlaubt es uns bei Beginn der Arbeit einfach den Abschnitt des Anderen aus zu kommentieren und unkompliziert dort weiter zu arbeiten wo der Andere aufgehört hat.

Kollaboratives Arbeiten

Die Komplexität und Grösse der Aufgabenstellung zwang uns immer mehr zu einer flexiblen Arbeitsweise und weg von einer groben Aufgabentrennung. Wir verbrachten also den Grossteil des Programmierens damit zusammen an einem Computer oder über Videochat alles zu erarbeiten. Wir haben anfangs probiert eigene

Versionen des Codes zu schreiben und dann zu vergleichen, kamen aber ziemlich schnell zum Schluss dass das keinen Sinn macht, da unser Code praktisch immer gleich war, da wir uns auf die gleichen Resourcen stützten und bei gleichen Quellen Hilfe bezogen.

Eigene Einflussfaktoren

Wir haben uns auf Einflussfaktoren beschränkt, die die Lebenszufriedenheit negativ beeinflussen, da unsere Erfahrungen mit der Pandemie mehrheitlich negativ waren und entschieden uns schlussendlich für die Variablen "P19L01", 'H19H27' & 'H19H23'. Sie stammen alle aus dem Datensatz vom Jahr 2019. Die erste Variable beantwortet die Frage ob die befragte Person Krankheit oder einen Unfall erlebt hat in diesem Jahr. Die zweite und dritte beziehen sich auf die Unterkunft und zwar ob es entweder Probleme damit gab (dreckig etc) oder ob sie zu klein war. Unsere Hypothese lautet demnach, dass Befragte, welche bei diesen Variablen ein "Ja" vermerkt haben, dementsprechend eine Verschlechterung der Lebenszufriedenheit erlebt haben.

Datenaufbereitung

Vorbereitung & Einlesen der Daten

Nach Download der Daten mussten wir die benötigten Variablen im Datensatz finden und selektieren. Wir nutzten dafür das Cheatsheet, das uns zur Verfügung gestellt wurde. Somit wussten wir bereits welche Datensätze nutzvoll sind für unsere Arbeit und welche nicht, und konnten diese sofort einlesen. (Zeilen 14-37)

Selektieren

Wir entschieden uns dazu in einem ersten Schritt alle Spalten, die uns interessieren, zu selektieren und jeweils in einem eigenen separaten Datensatz zu speichern. (Zeilen 41-72)

Zusammenführen der Daten

Der zweite Schritt bestand darin die neuen kleinen Datansätze zu einem grossen zusammen zu mergen. Unser Datensatz "Covid2020" beinhaltet die Angaben zu den Personen, welche an der Covidstudie im Jahr 2020 mitgemacht haben. Uns interessieren lediglich die Daten derer, die da drin vorkommen. Da jeder Teilnehmer dieser Studien eine eindeutige Identifikationsnummer (Variable IDPERS) besitzt, entschieden wir uns damit jeweils alle Datensätze mit einem left join zu mergen, beginnend mit dem "Covid2020" Datensatz. Der einzige Datensatz, welcher nicht über die Variable "IDPERS" zusammengeführt werden konnte, war "H19" (Zeilen 74-95). Die Daten stammen aus Befragungen zu den Haushälten als Ganzes und nicht den einzelnen Personen. Hier geschieht der merge über die Variable "IDHOUS19". Die folgende Abbildung ist eine Skizze des Zusammenführens der Daten. Jeder Strich zwischen den Tabellen steht dabei für einen left join. Das Endresultat ist die letzte Tabelle mit dem Titel "data", welche unseren finalen Datensatz darstellt. An diesem Punkt besteht er aus 5843 Zeilen und 27 Spalten.

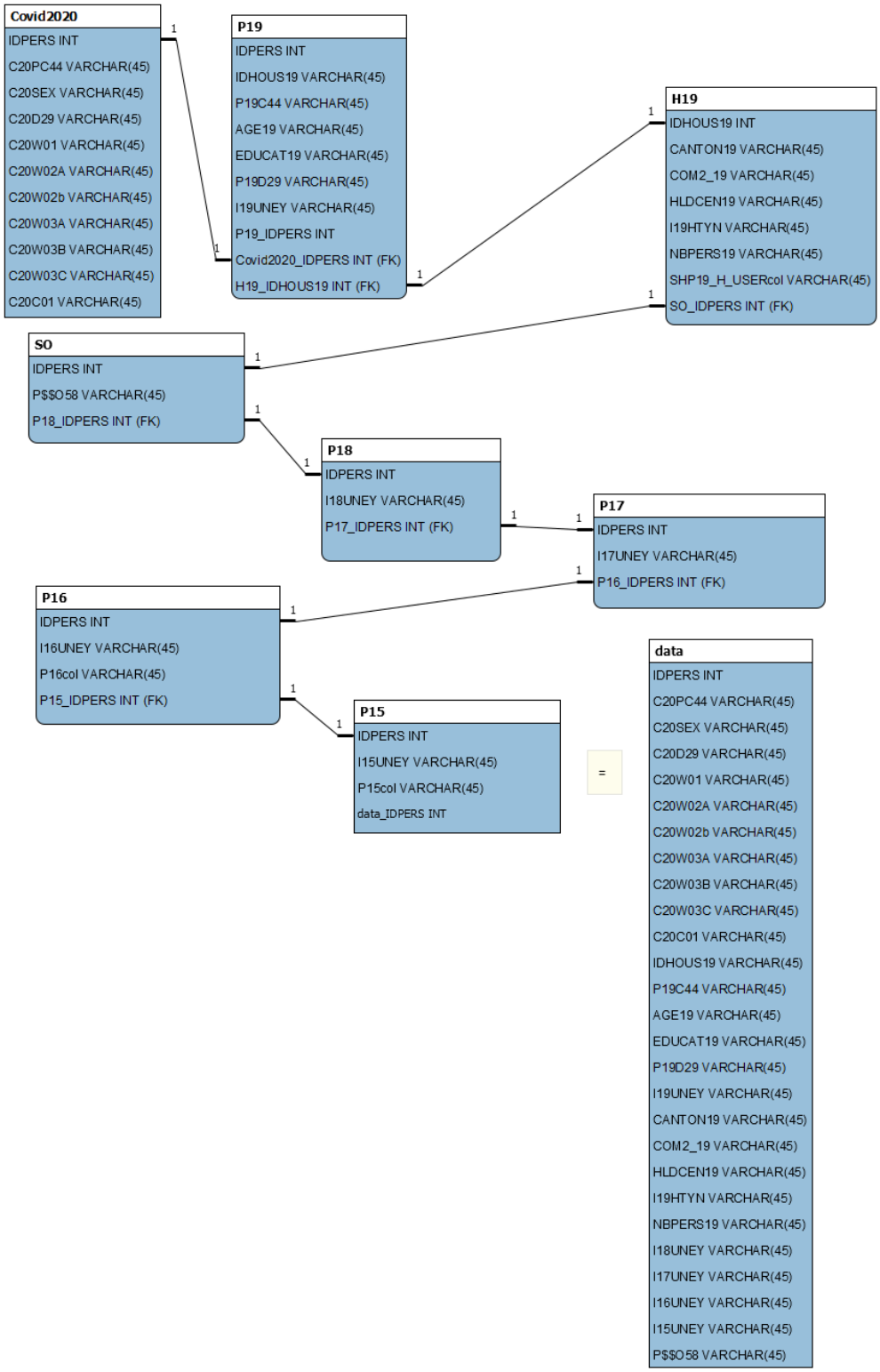

Figure 1: Skizze zum Zusammenführen der Daten

Bereinigen der Daten

Der Datensatz enthielt viele Missing Value Codes, also fehlerhafte oder fehlende Einträge. Sie wurden als Werte unter 0 codiert. Als erstes ersetzten wir diese Werte durch Na, damit wir bei den späteren Darstellungen diese Werte einfacher ignorieren können und möglichst wenig Informationsverlust haben. (Zeilen 102-104) Zusätzlich mussten wir die Spalten umbenennen, zu Namen die für uns sinnvoller erschienen, als Zahlen codierte Werte umwandeln (1 -> "in Beziehung" bspw.) und neue Spalten aus alten erstellen. Die genauen Arbeitsschritte finden Sie im Code als Kommentare. Schlussendlich löschten wir alle Spalten die nicht mehr benötigt wurden. Nach diesem Schritt bestand der Datensatz aus 5630 Zeilen und 16 Spalten.

3 zusätzliche Variablen hinzufügen

Die unter "Eigene Einflussfaktoren" beschriebenen Variablen wurden in diesem Schritt mit dem Hauptdatensatz, ebenfalls über einen left join, zusammengeführt. Dies bringt uns zur Enddimension von 5630 Zeilen und 19 Spalten.

Prüfen der Datenqualität

Wie bereits im Abschnitt "Bereinigen der Daten" erwähnt wurde, wandelten wir die Missing Value Codes zu Nas um, da man besser damit arbeiten kann. Wir wollen nun herausfinden, wieviele fehlerhafte oder fehlende Daten unsere Variablen aufweisen. Wir können uns nun einfach die Anzahl Nas pro Variable zählen und ausgeben lassen. (Zeile 238) Das Resultat wird in der folgenden Tabelle veranschaulicht:

Table 1: Missing Values pro Variable

Variable	Anzahl
IDPERS	0
IDHOUS19	0
Alter	0
Höchstes_Ausbildungszertifikat	0
Kanton	0
Gemeindetyp	0
Haushaltstyp	3
Geschlecht	0
Beziehungsstatus	72
Beziehungsstatusänderung	75
Kurzfristige Änderung Arbeitssituation	0
Covid_infektion	99
EinkommenProKopf	383
Finanzielle Probleme in Jugend	352
Arbeitslos_15_19	0
LZ_änderung	63
Unterkunftsprobleme	11
Unterkunft zu klein	8
Erkrankung,Unfall	4

Wir können sehen, dass bei gewissen Variablen manche Leute keine Abgaben machen konnten oder wollten wie beispielsweise bei dem Beziehungsstatus. Da wir Werte wie "Beziehungsstatusänderung" aus zwei Variablen errechnen (Zeile 147-152), nämlich C20D29 (Beziehungsstatus im Jahr 2020) und P19D29 (Beziehungsstatus im Jahr 2019), erhalten wir so 75 Missing Values, weil die erste Variable 72 und die zweite

3 fehlende Angaben hat. Auffallend sind vorallem die Variablen “Einkommen pro Kopf” und “Finanzielle Probleme in Jugend”. Hier finden wir die höchste Anzahl an Missing Values. Es fehlt im Schnitt zu jeder vierzehnten Person eine Angabe und sollte womöglich nicht berücksichtigt werden bei der Auswertung.

Plausibilität der Daten

Zur Überprüfung der Plausibilität unserer Daten haben wir beim Bundesamt für Statistik einen Datensatz zur Gesamtpopulation der Schweiz runtergeladen. Er enthält unteranderem Angaben zu den verschiedenen Altersgruppen und ihrem Anteil an der Bevölkerung, den verschiedenen Kantonen und wieviel Leute dort wohnen und der prozentuale Anteil an Frauen und Männern in der Schweiz. Um herauszufinden wie akkurat unser Datensatz die schweizer Bevölkerung wiederspiegelt stellen wir diese Messgrössen von unserem Datensatz zusammen mit denen vom Bundesamt für Statistik in einer Grafik dar.

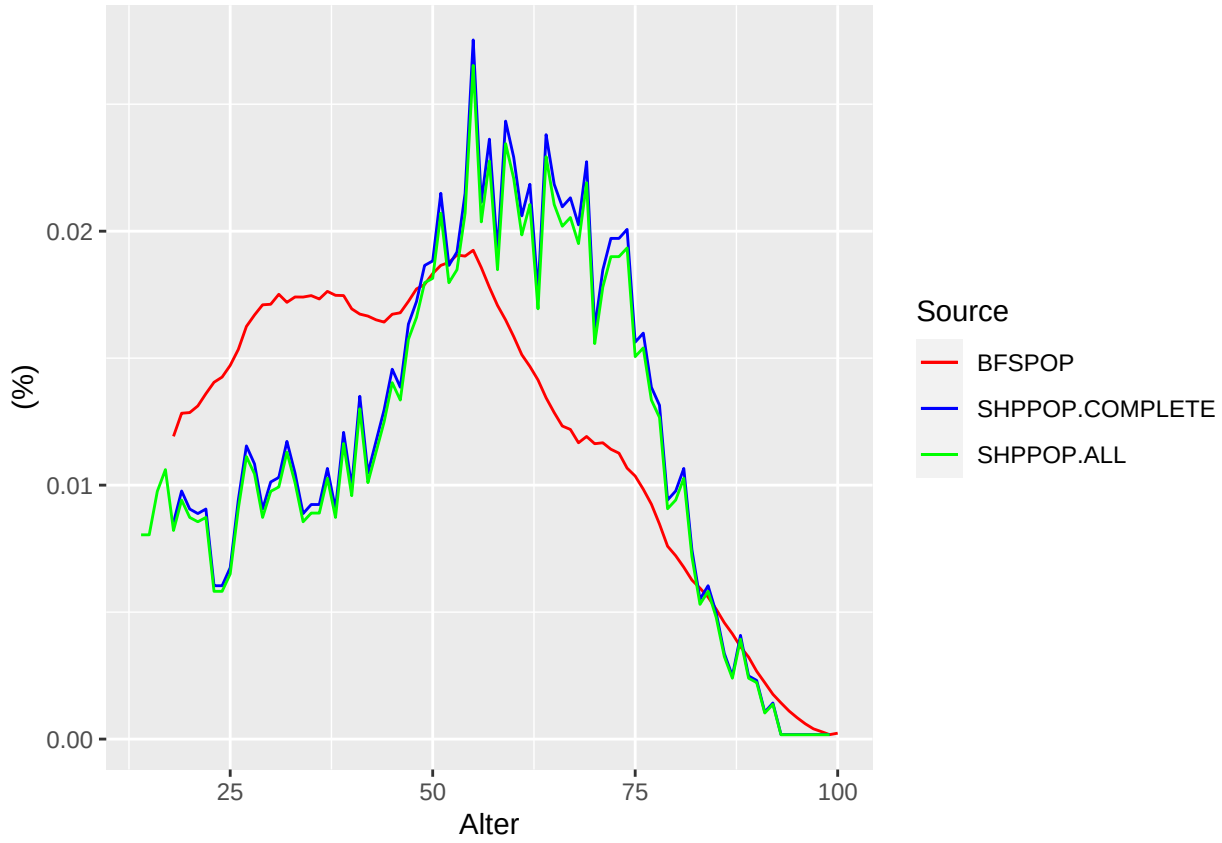

Figure 2: Plausibilität nach Alter

In der Grafik sehen wir die verschiedenen Altersgruppen und den Anteil den sie an der Population ausmachen (rot) und wie sie bei unserem Datensatz vor (grün) und nach (blau) dem Filtern vertreten sind. Wie man sieht spiegeln unsere Daten die Bevölkerung ziemlich schlecht wieder. Leute zwischen 18-50 Jahre alt sind unterrepräsentiert, während die zwischen 50 und 80 stark überrepräsentiert sind.

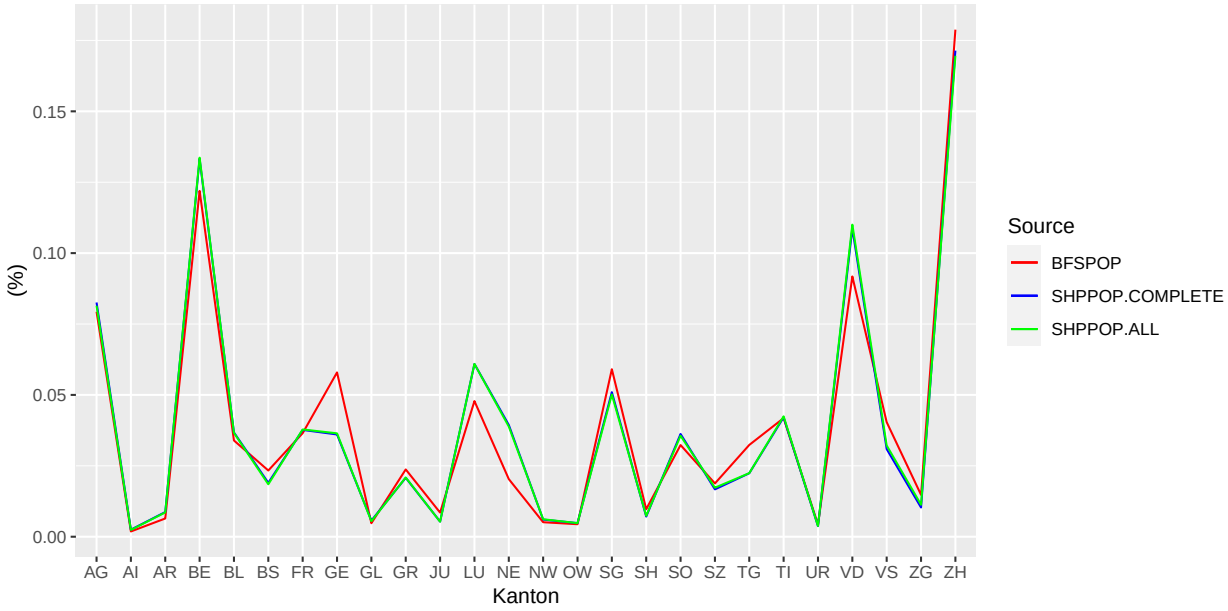

Figure 3: Plausibilität nach Kanton

Erkennbar hier ist, dass alle Kantone gut wiederspiegelt sind. Die einzigen zwei Ausreisser sind Genf und Luzern, wobei die Unterschiede gering sind.

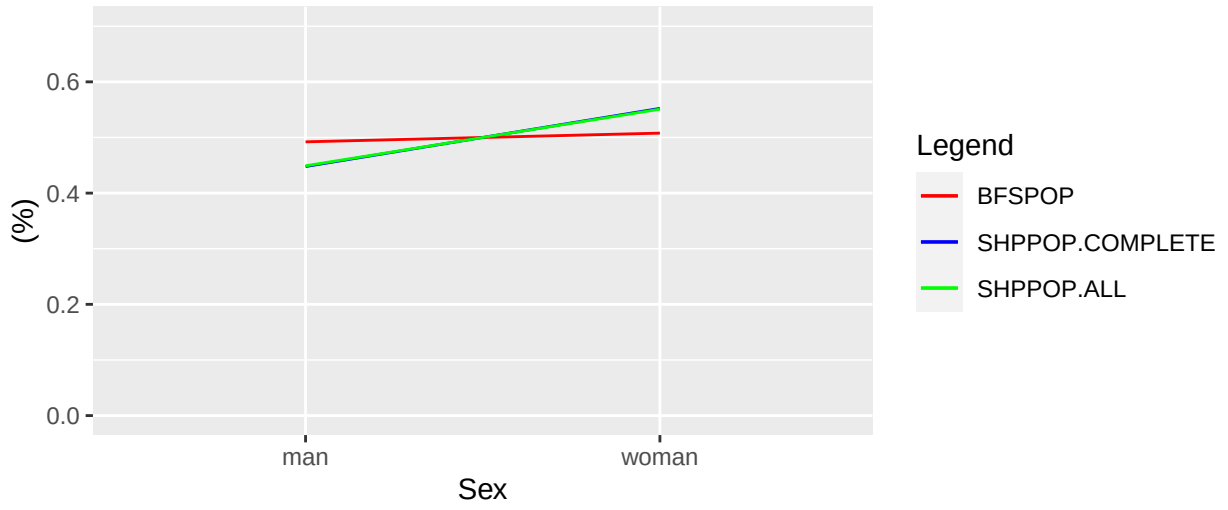

Figure 4: Plausibilität nach Geschlecht

Wie man sieht sind die Geschlechter auch nicht exakt wiederspiegelt. Frauen sind überrepräsentiert. Wir können also zusammenfassend sagen, dass wir ein Übermass an Frauen und älteren Leuten in unserem Datensatz haben. Der Datensatz ist also nicht sehr represäntativ in Bezug auf die normale Bevölkerung der Schweiz.

Auswertung

Univariate Analysen

Um uns einen Überblick über die möglichen Einflussfaktoren zu verschaffen hier zunächst univariate Darstellungen zu den Variablen:

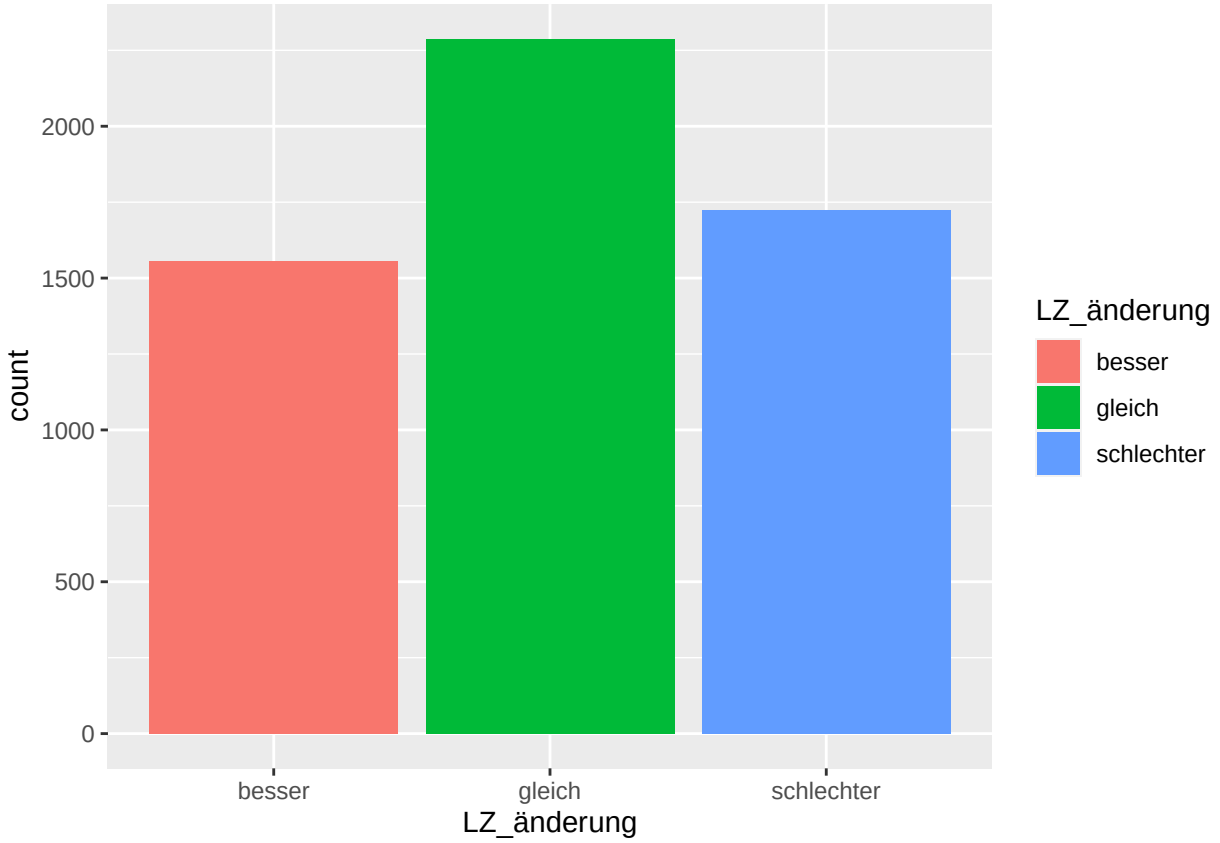

Figure 5: univariate Darstellung zur Lebenszufriedenheitsänderung

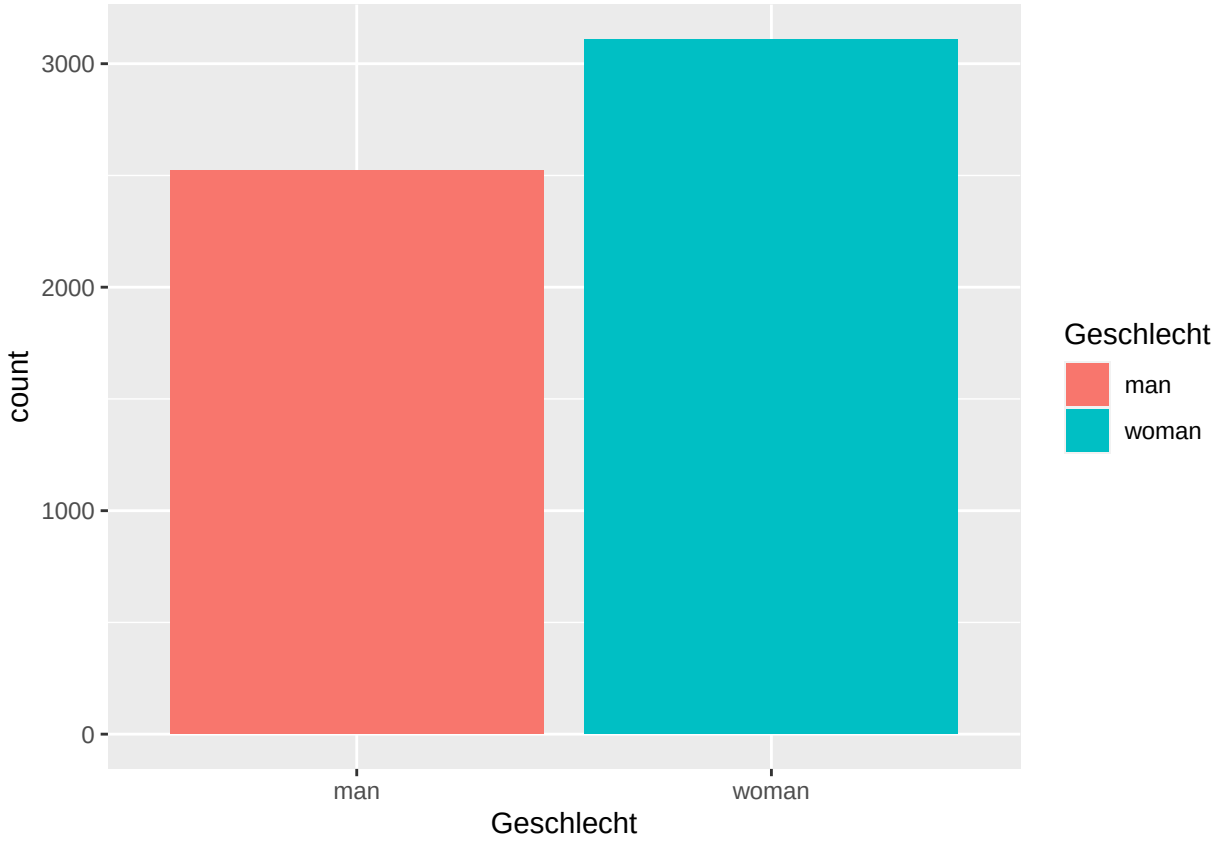

Figure 6: univariate Darstellung zum Geschlecht

Hier ist nochmals ersichtlich, dass Frauen überrepräsentiert sind.

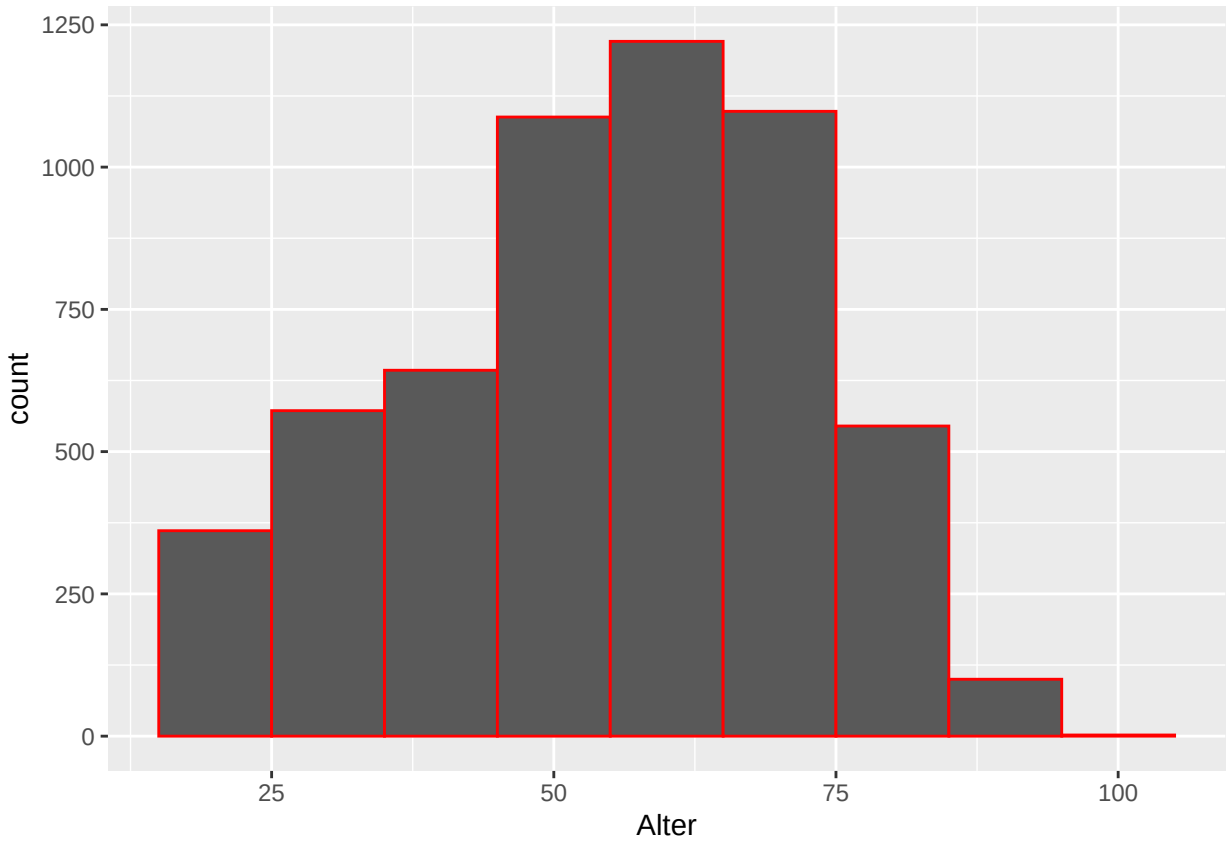

Figure 7: univariate Darstellung zum Alter

Hier ist nochmals ersichtlich, dass Personen im Alter von 50-80 überrepräsentiert sind.

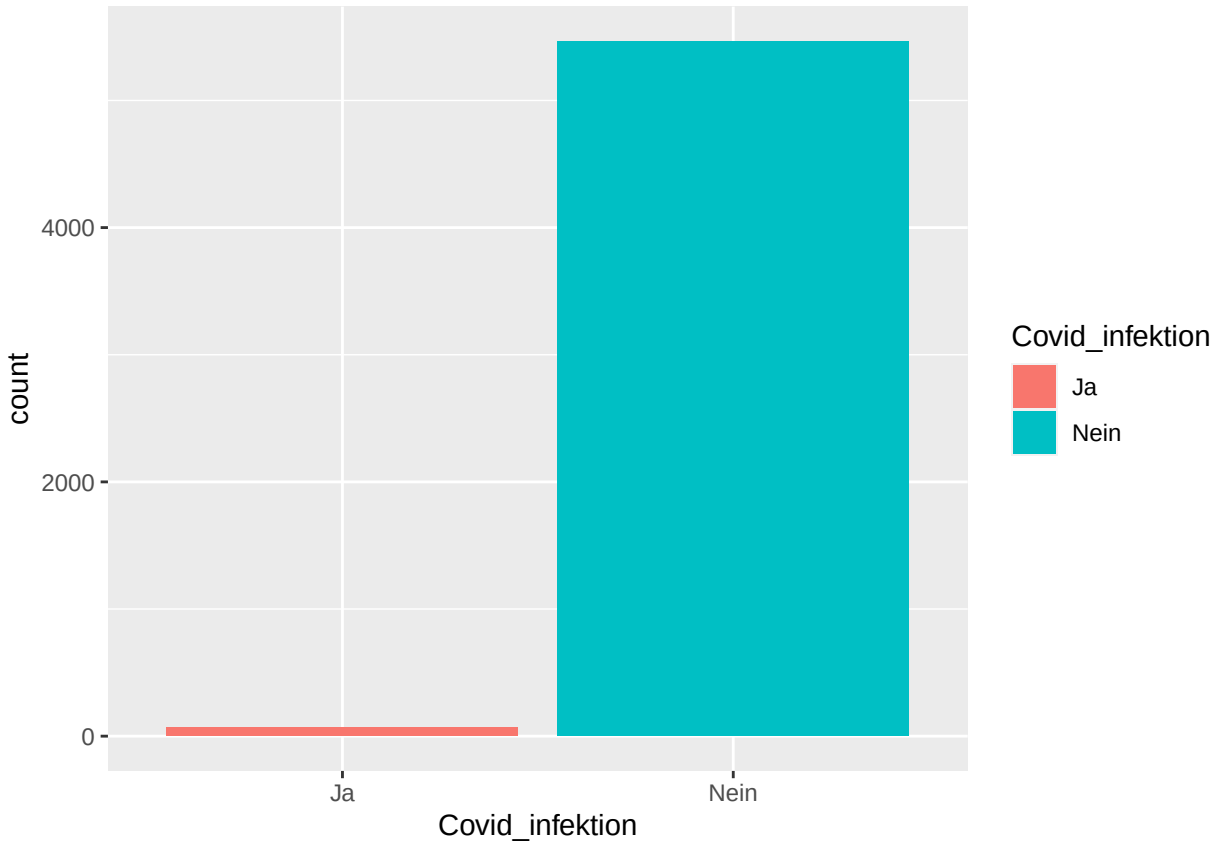

Figure 8: univariate Darstellung zur Covid Infektion

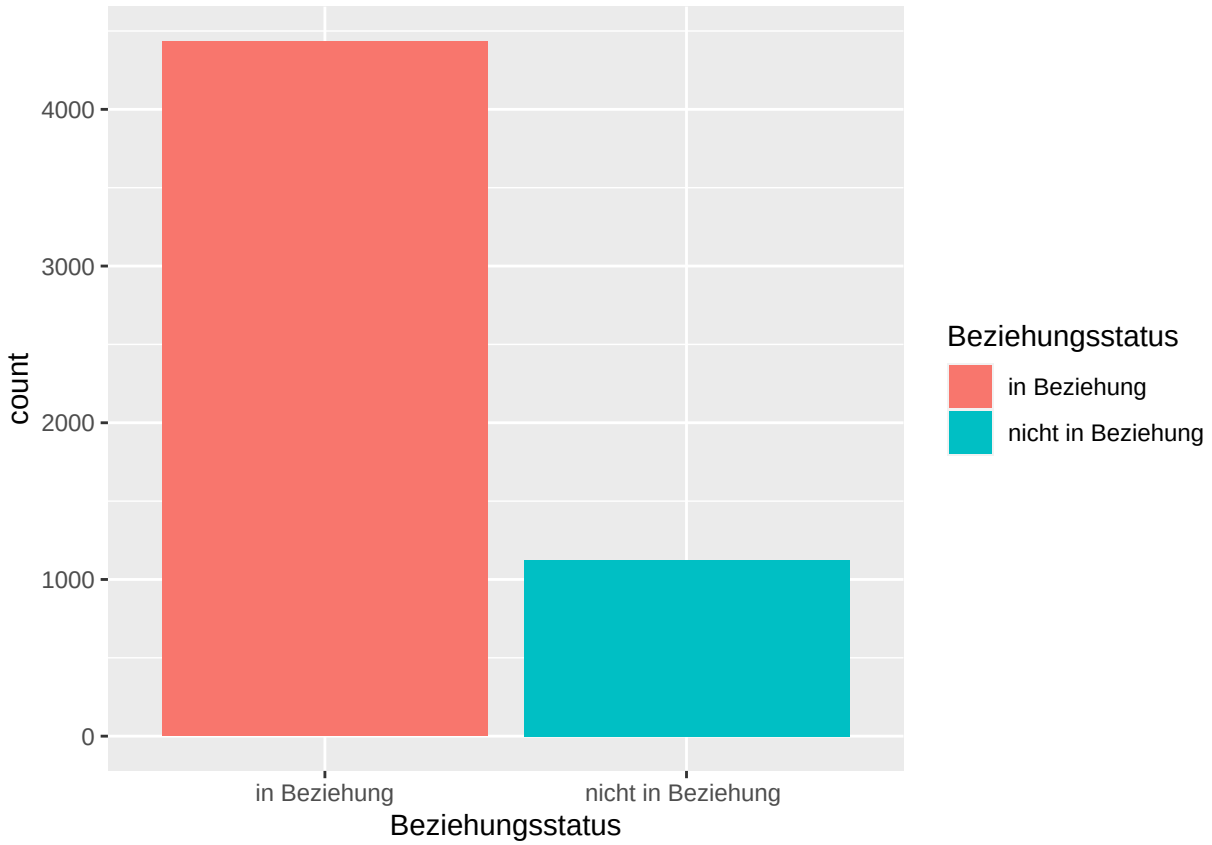

Figure 9: univariate Darstellung zum Beziehungsstatus

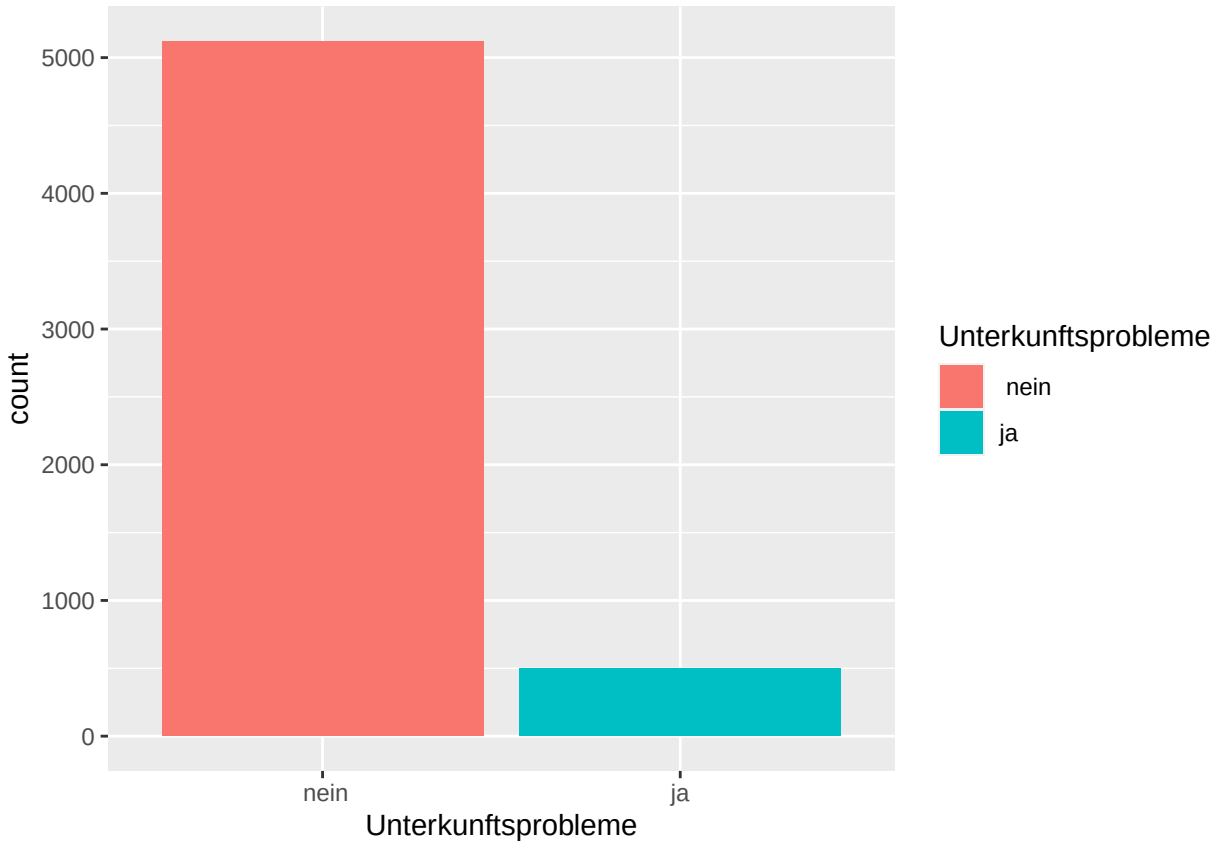

Figure 10: univariate Darstellung zu den Unterkunftsproblemen

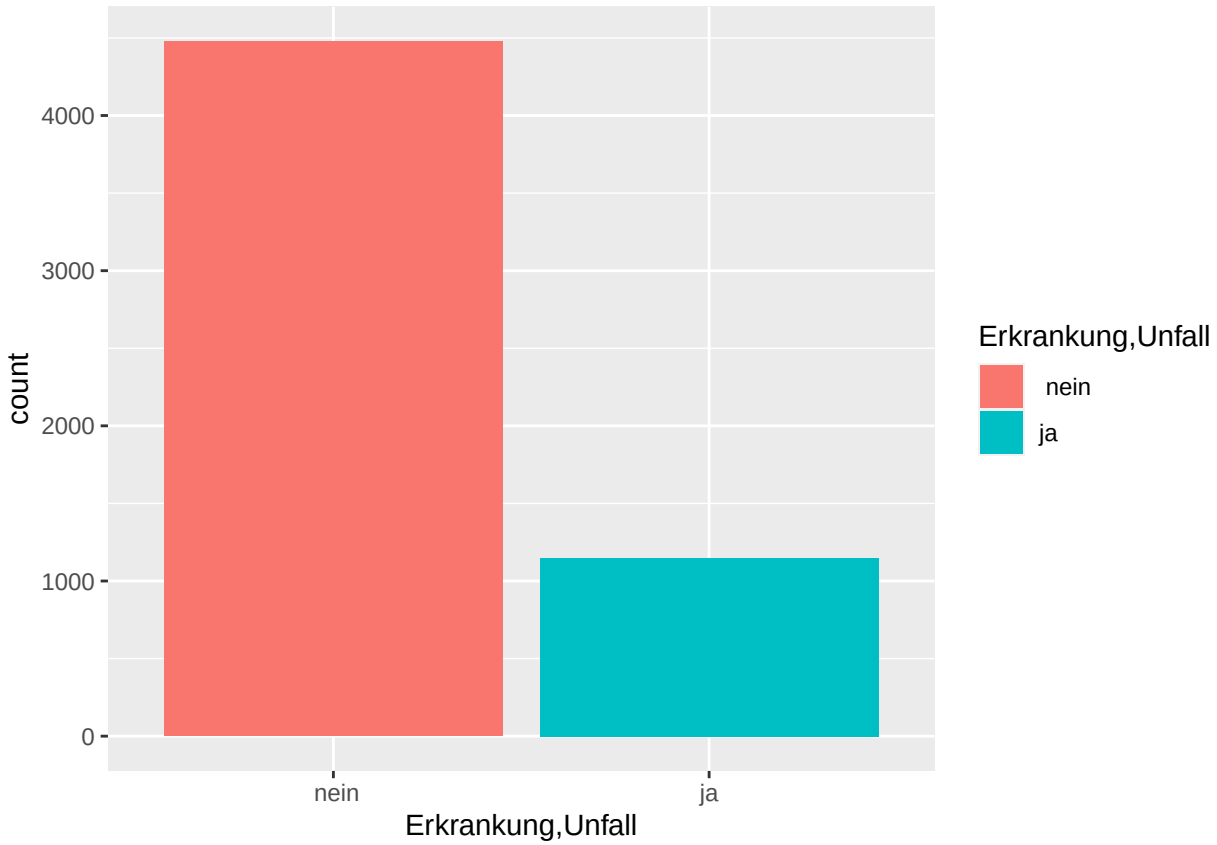

Figure 11: univariate Darstellung zu Erkrankung, Unfall

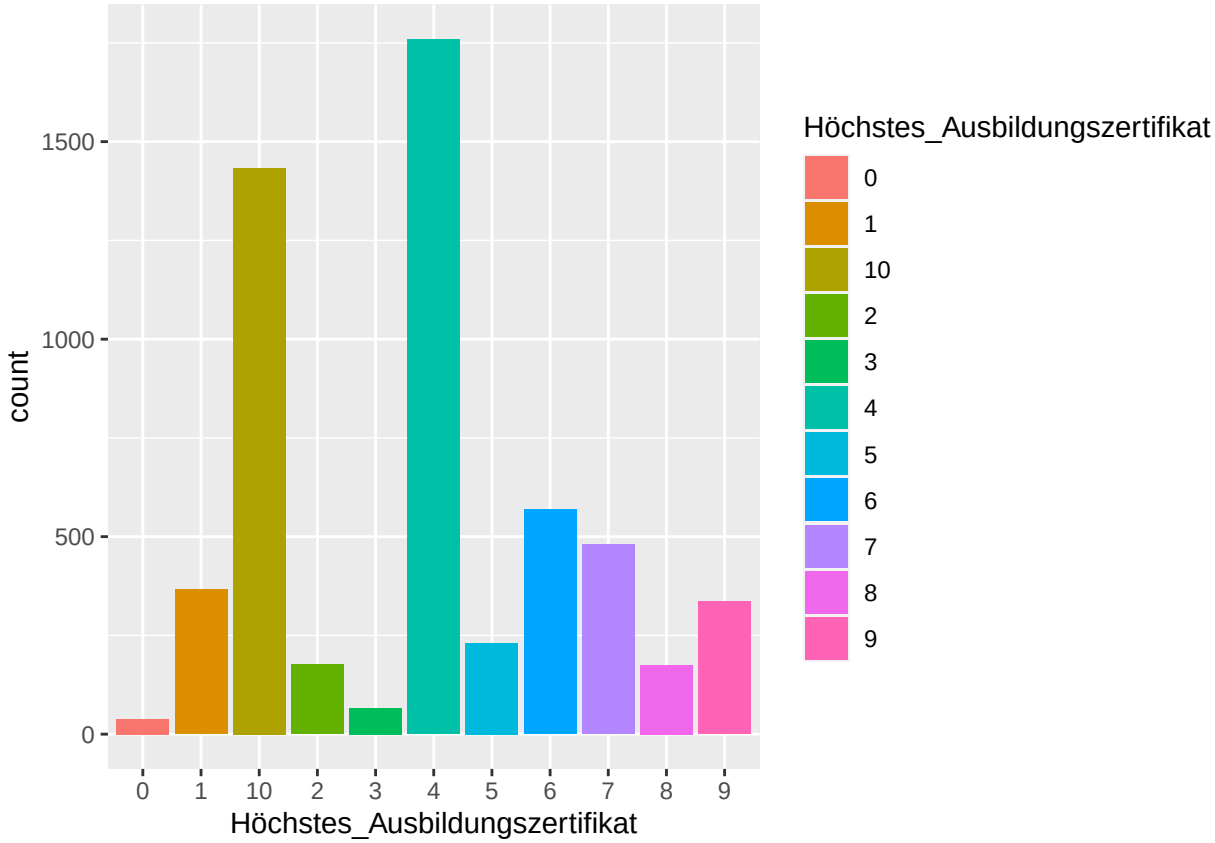

Figure 12: univariate Darstellung zum Höchsten Ausbildungszertifikat

Aus Platzgründen hier die Beschriftungen für die Werte in der Legende: 0 incomplete compulsory school 1 compulsory school, elementary vocational training 2 domestic science course, 1 year school of commerce 3 general training school 4 apprenticeship (CFC, EFZ) 5 full-time vocational school 6 bachelor/maturity 7 vocational high school with master certificate, federal certificate 8 technical or vocational school 9 vocational high school ETS, HTL etc. 10 university, academic high school, HEP, PH, HES, FH

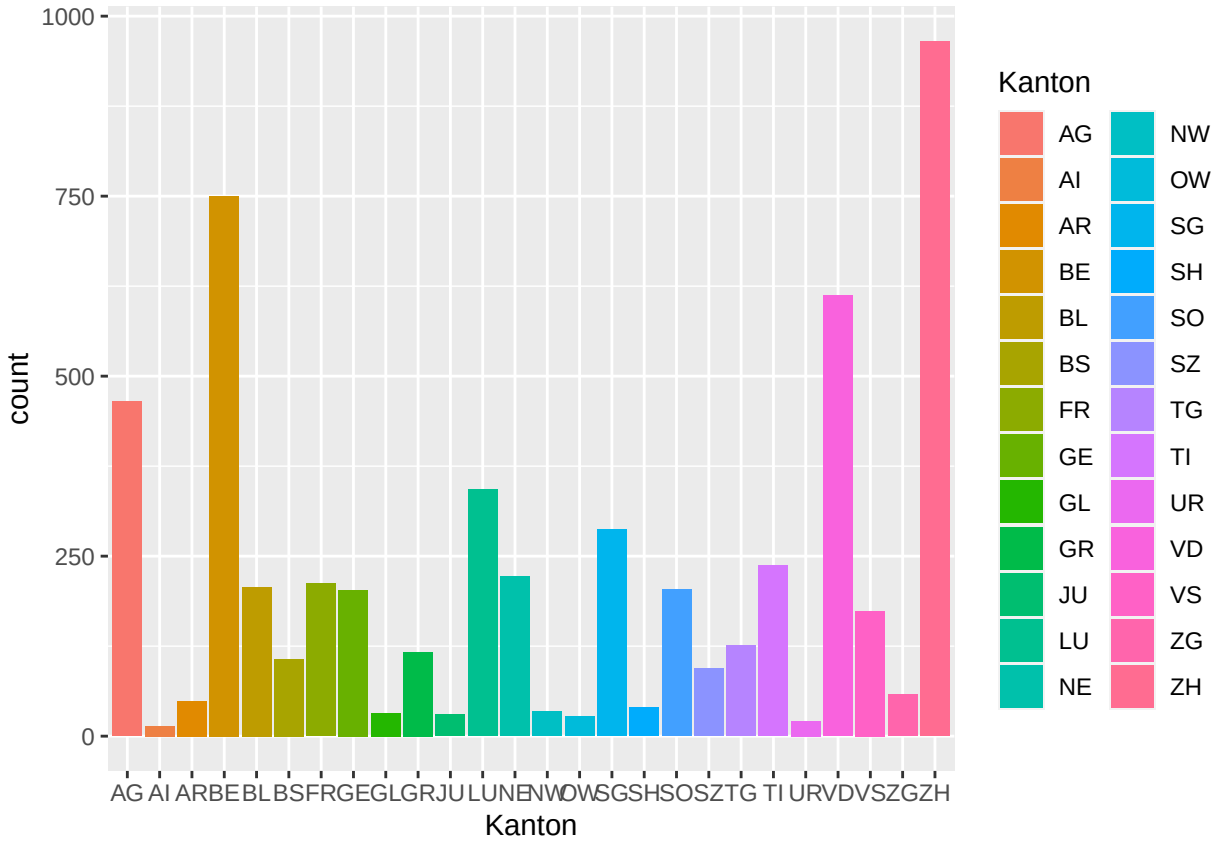

Figure 13: univariate Darstellung zu den Kantonen

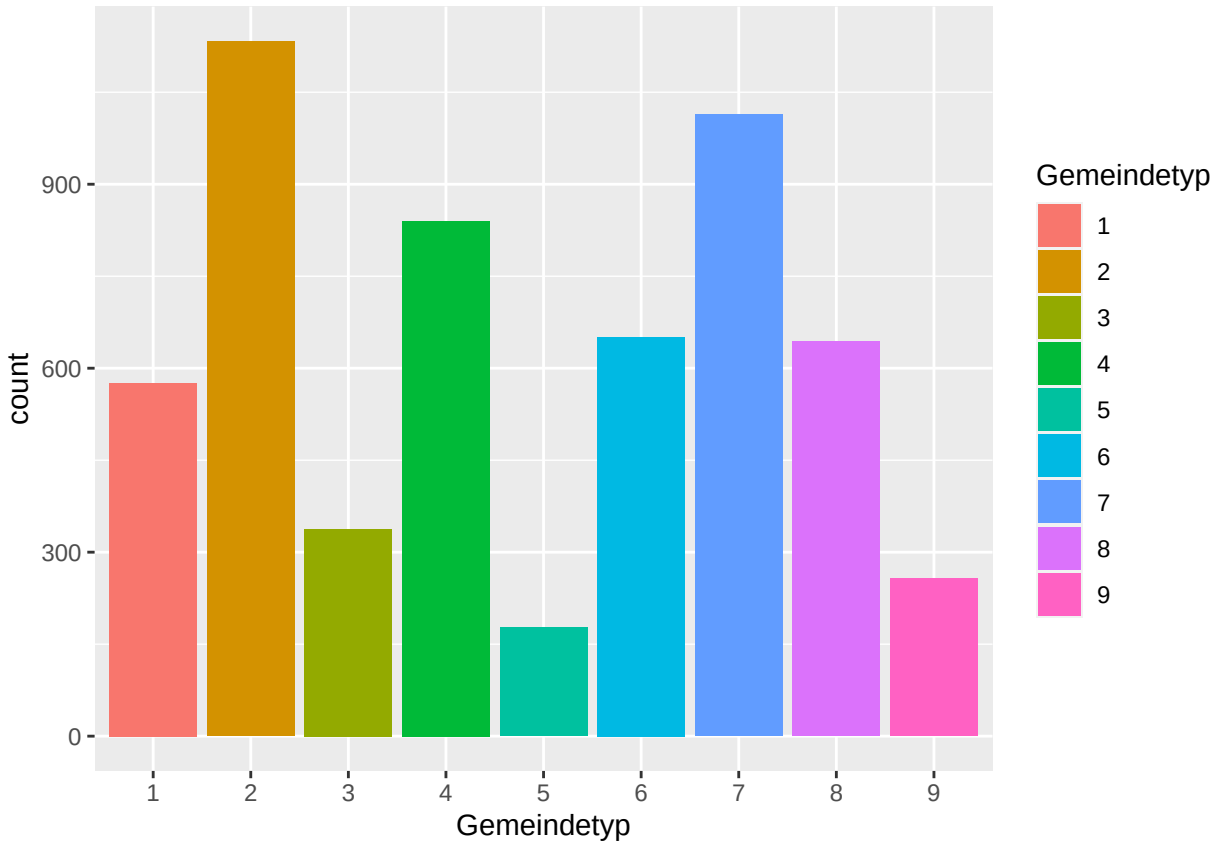

Figure 14: univariate Darstellung zum Gemeindetyp

Aus Platzgründen hier die Beschriftungen für die Werte in der Legende: 1 Centres 2 Suburban communes 3 Wealthy communes 4 Peripheral urban communes 5 Tourist communes 6 Industrial and tertiary sector communes 7 Rural commuter communes 8 Mixed agricultural communes 9 Peripheral agricultural communes

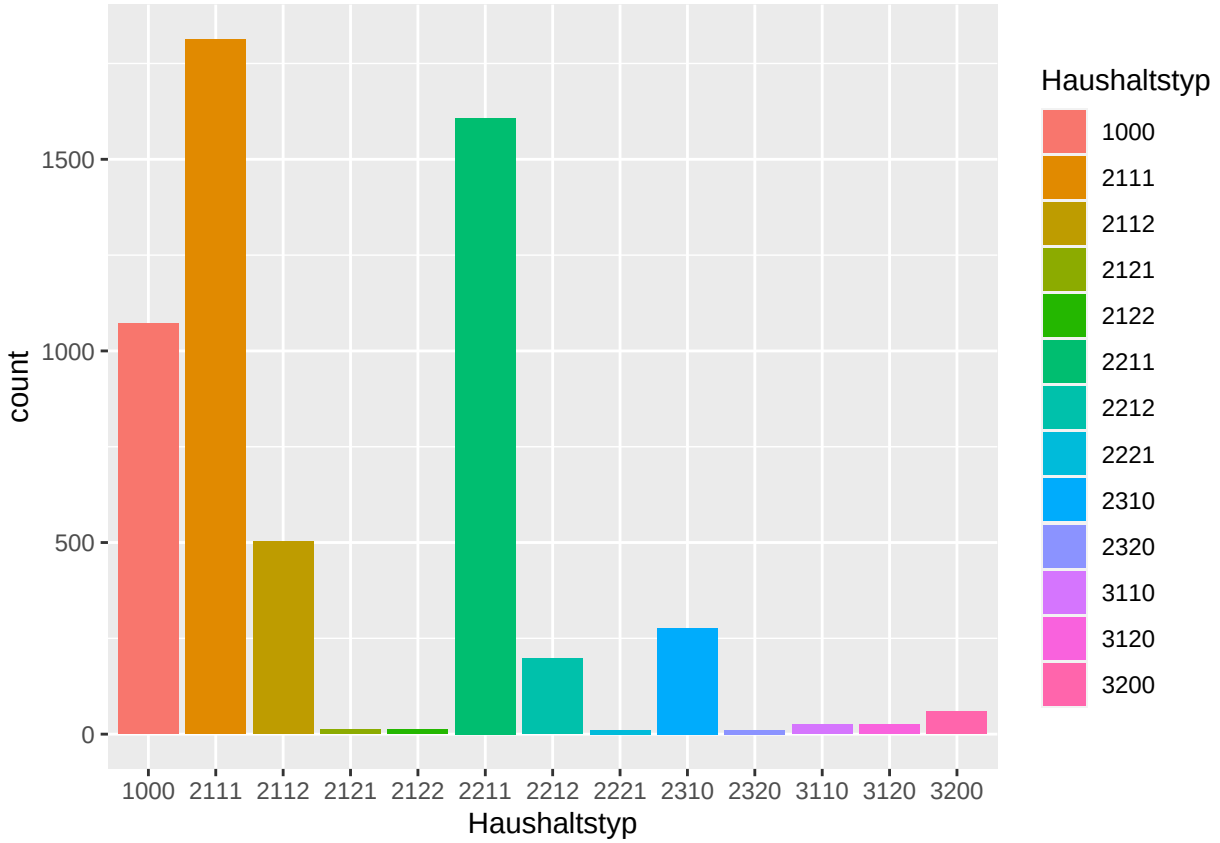

Figure 15: univariate Darstellung zum Haushaltstyp

Aus Platzgründen hier die Beschriftungen für die Werte in der Legende: 1000 One-person private households 2111 Married couple without children 2112 Consensual couple without children 2121 Married couple without children and another person 2122 Consensual couple without children and another person 2211 Married couple with children 2212 Consensual couple with children 2221 Married couple with children and another person 2222 Consensual couple with children and another person 2310 One parent with children 2320 One parent with children and another person 3110 Other types of households with only related family 3120 Other types of households with and without related family 3200 Other types of households without related family

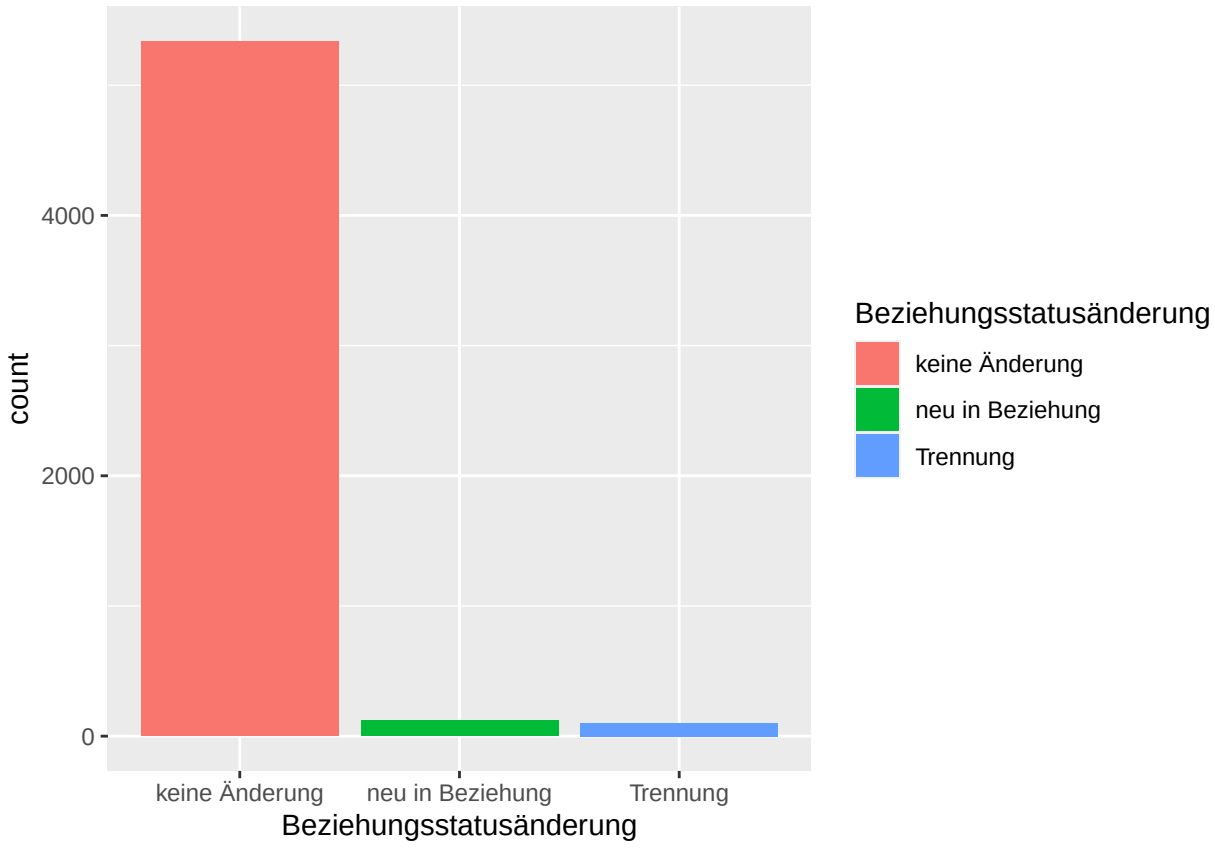

Figure 16: univariate Darstellung zur Beziehungsstatusänderung

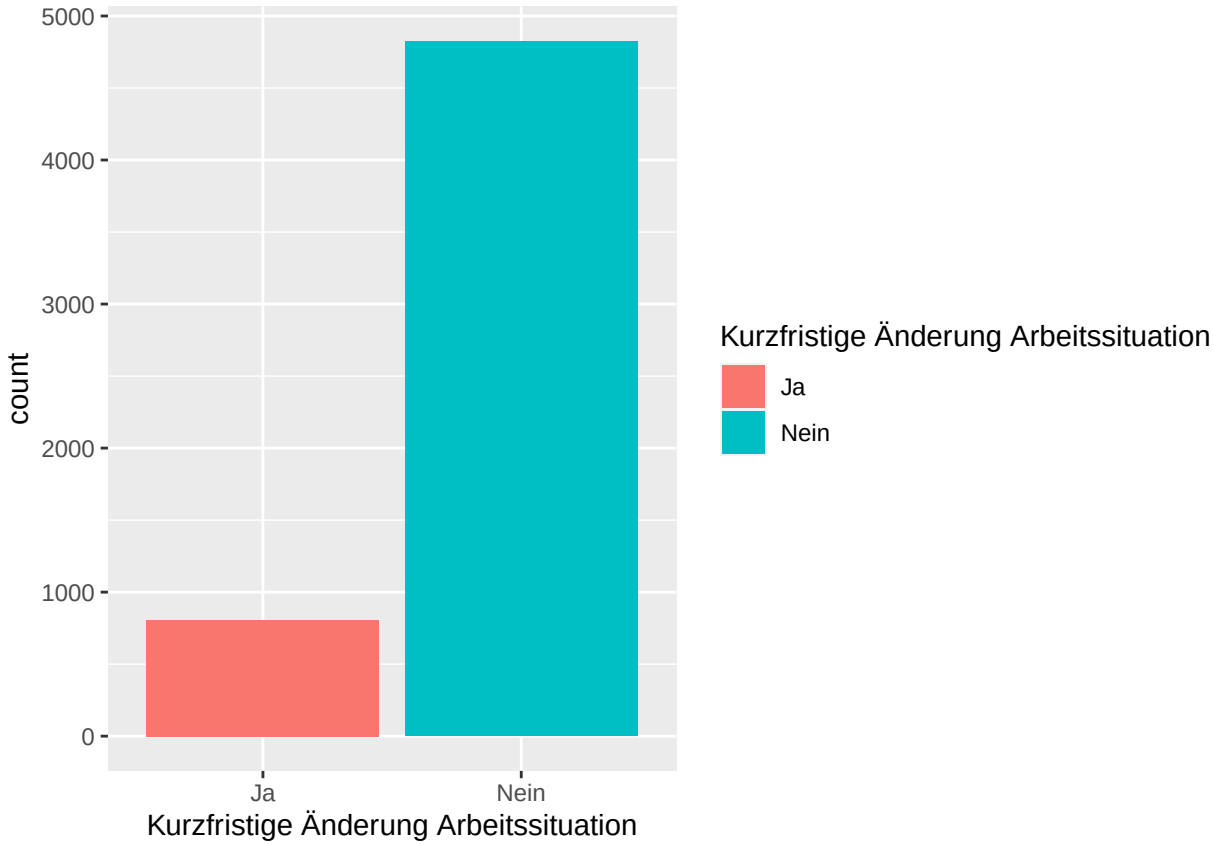

Figure 17: univariate Darstellung zur Kurzfristien Änderung der Arbeitssituation

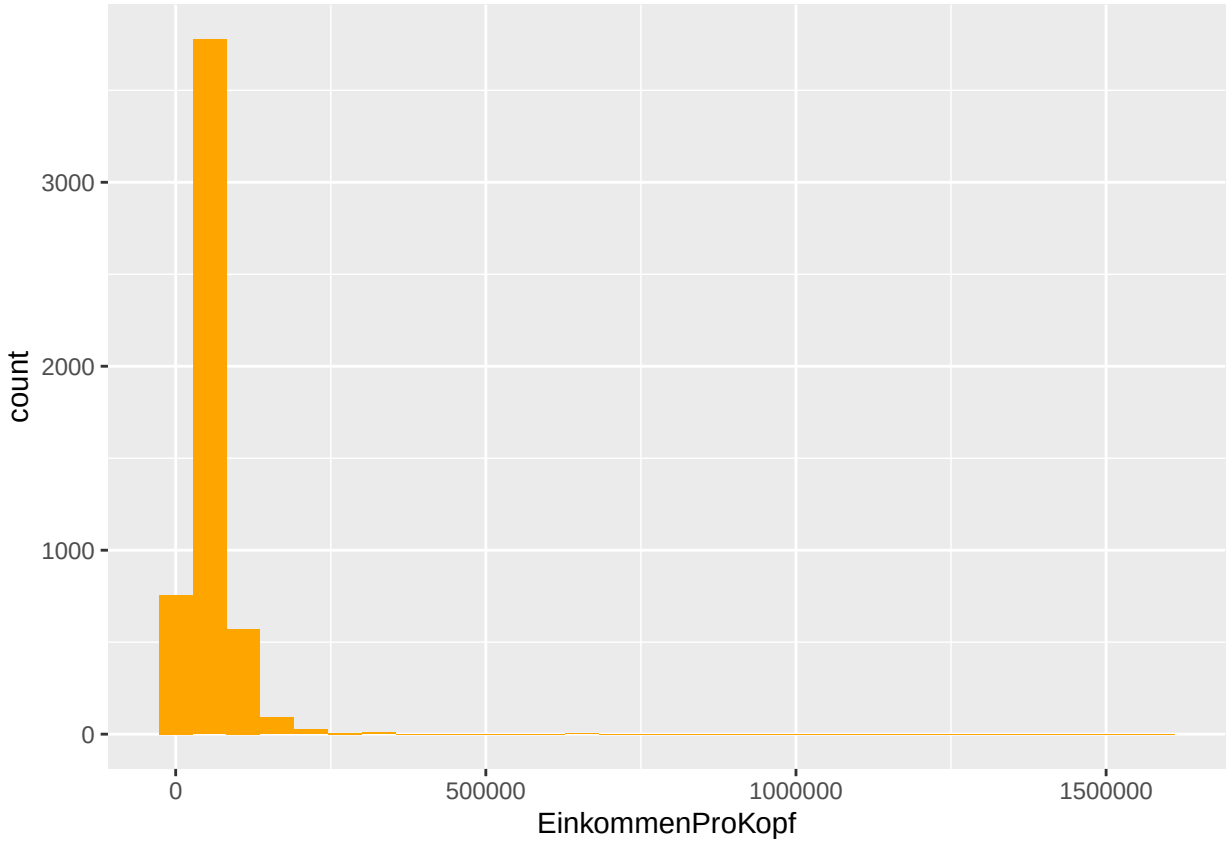

Figure 18: univariate Darstellung zum Einkommen pro Kopf

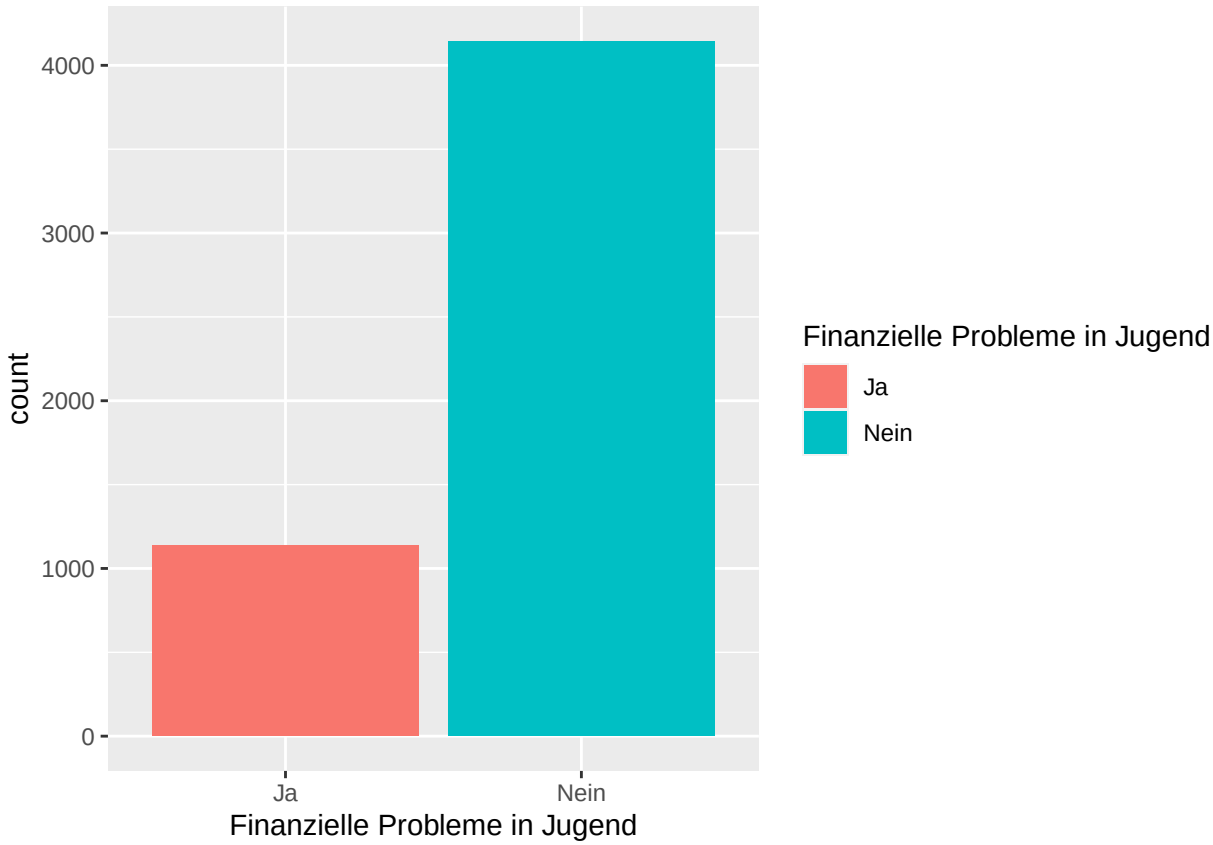

Figure 19: univariate Darstellung zu den Finanziellen Problemen in der Jugend

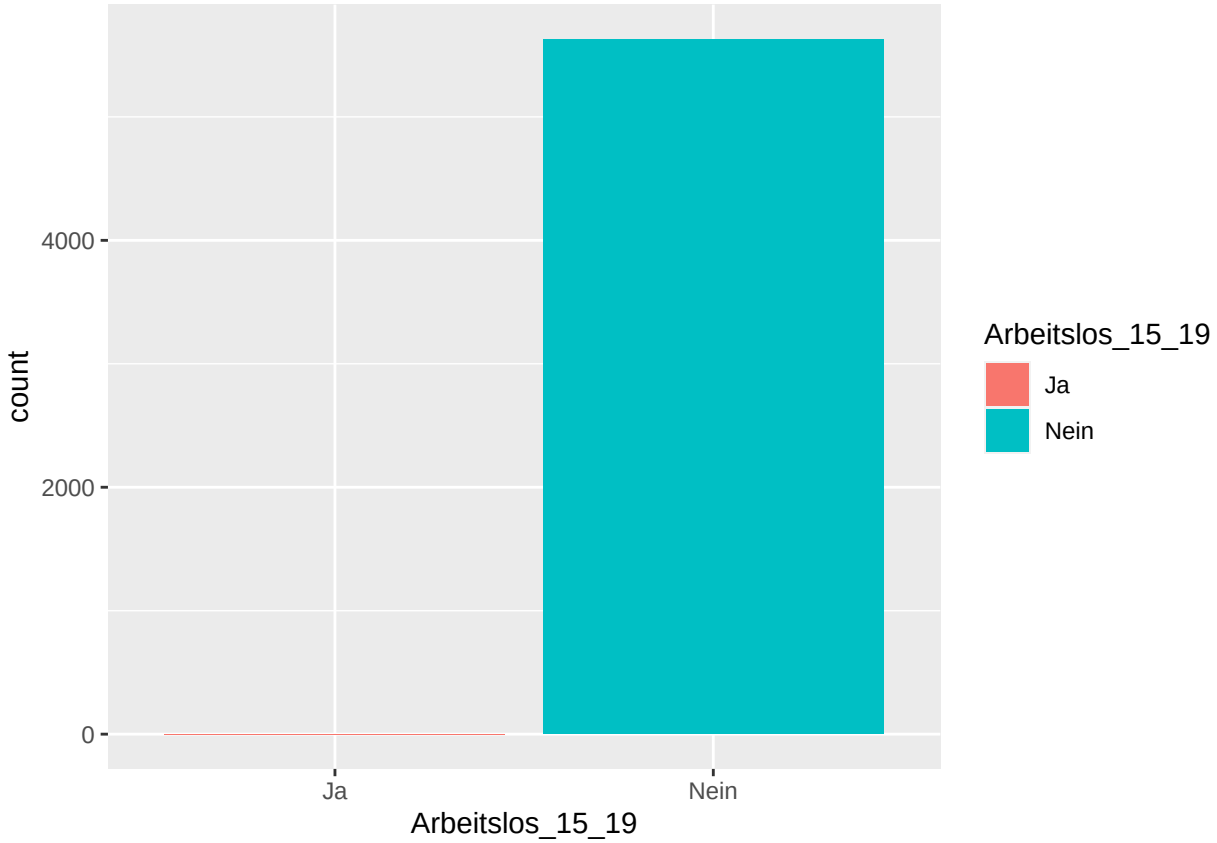

Figure 20: univariate Darstellung zur Arbeitlosigkeit

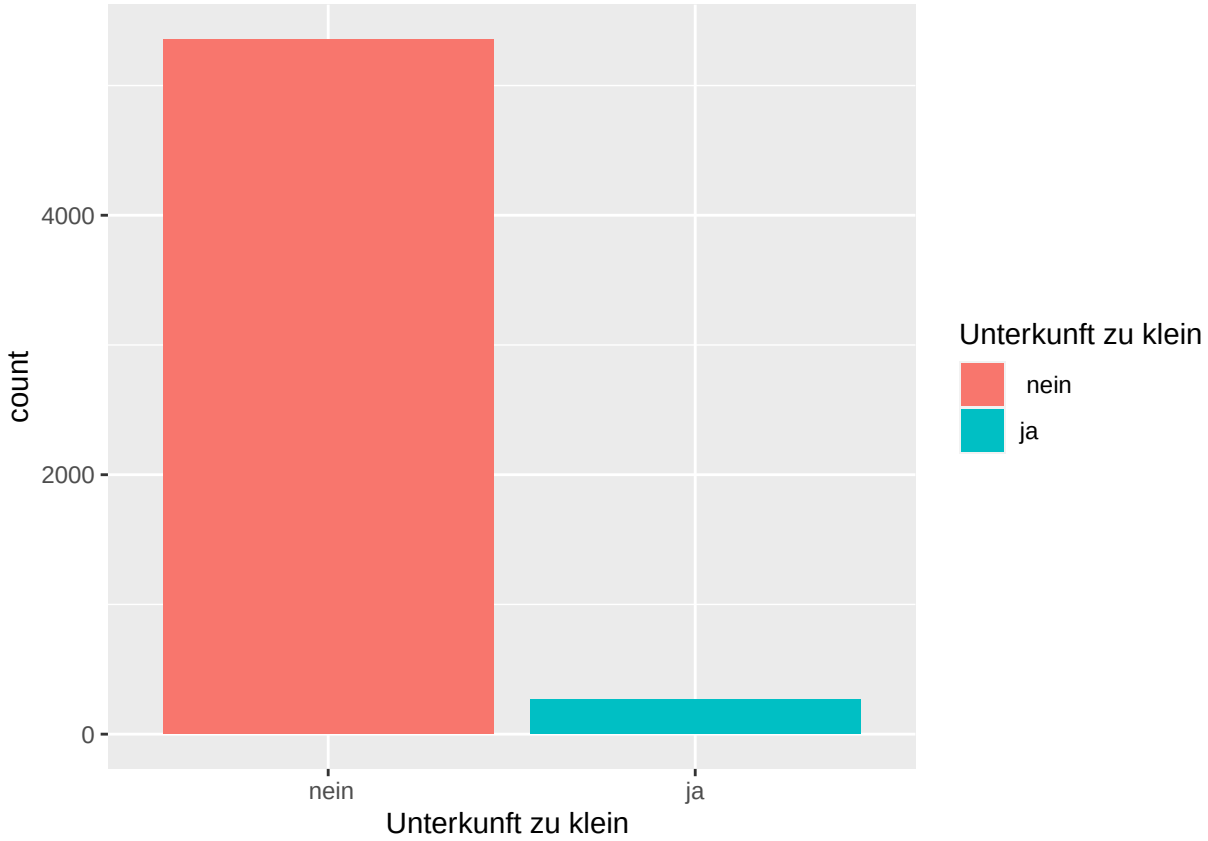

Figure 21: univariate Darstellung zur Grösse der Unterkunft

Bivariate Analysen

Um unsere Leitfrage beantworten zu können, stellen wir hier in einem nächsten Schritt jeweils Variablen, von denen wir denken, dass sie einen Einfluss auf die Lebenszufriedenheit haben könnten und die effektive Änderung der Lebenszufriedenheit dar.

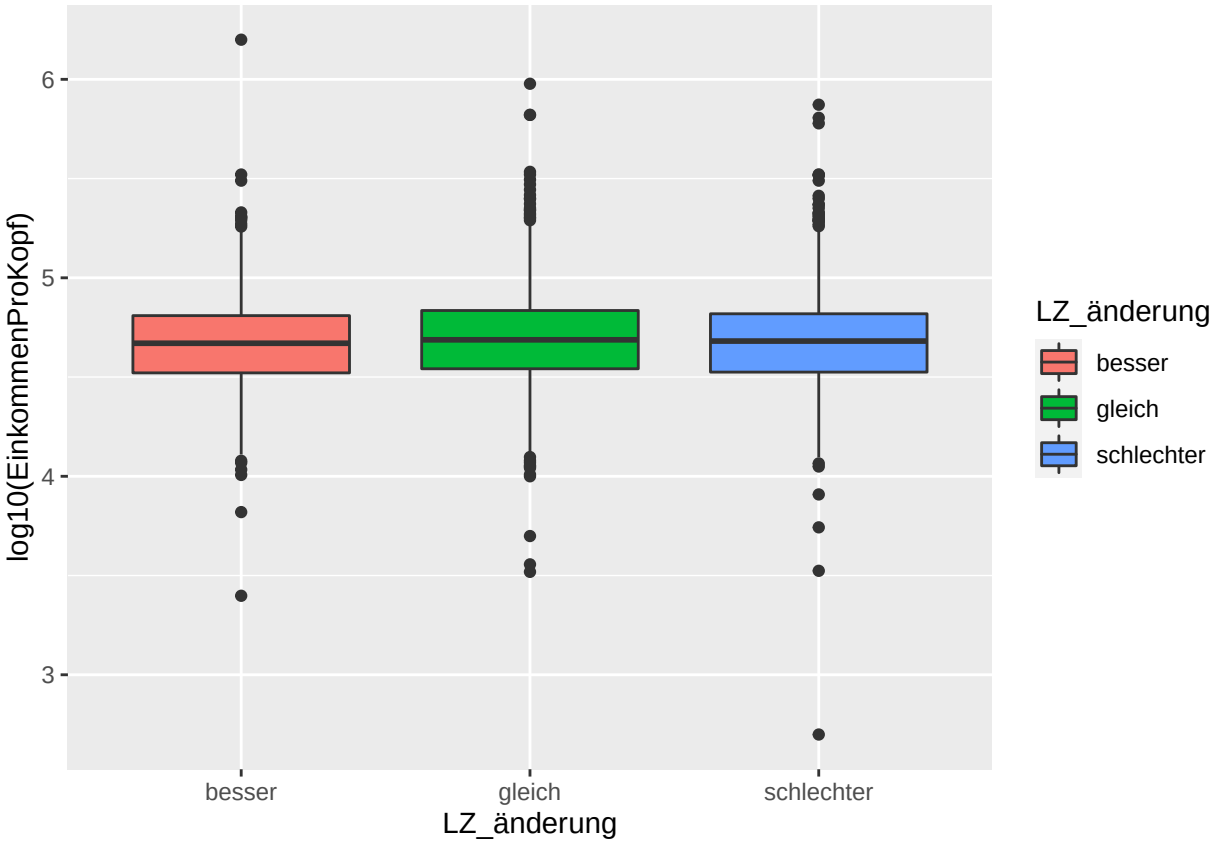

Figure 22: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf das Einkommen pro Kopf

Unsere Hypothese lautete, dass Personen mit geringerem Einkommen pro Kopf stärker unter der Pandemie litten, da der Virus einen grossen negativen finanziellen Einfluss auf die Bevölkerung hatte. Wie man der Grafik entnehmen kann, spiegelt unser Datensatz das jedoch nicht wieder. Der Median des Einkommens ist bei allen Gruppen etwa gleich.

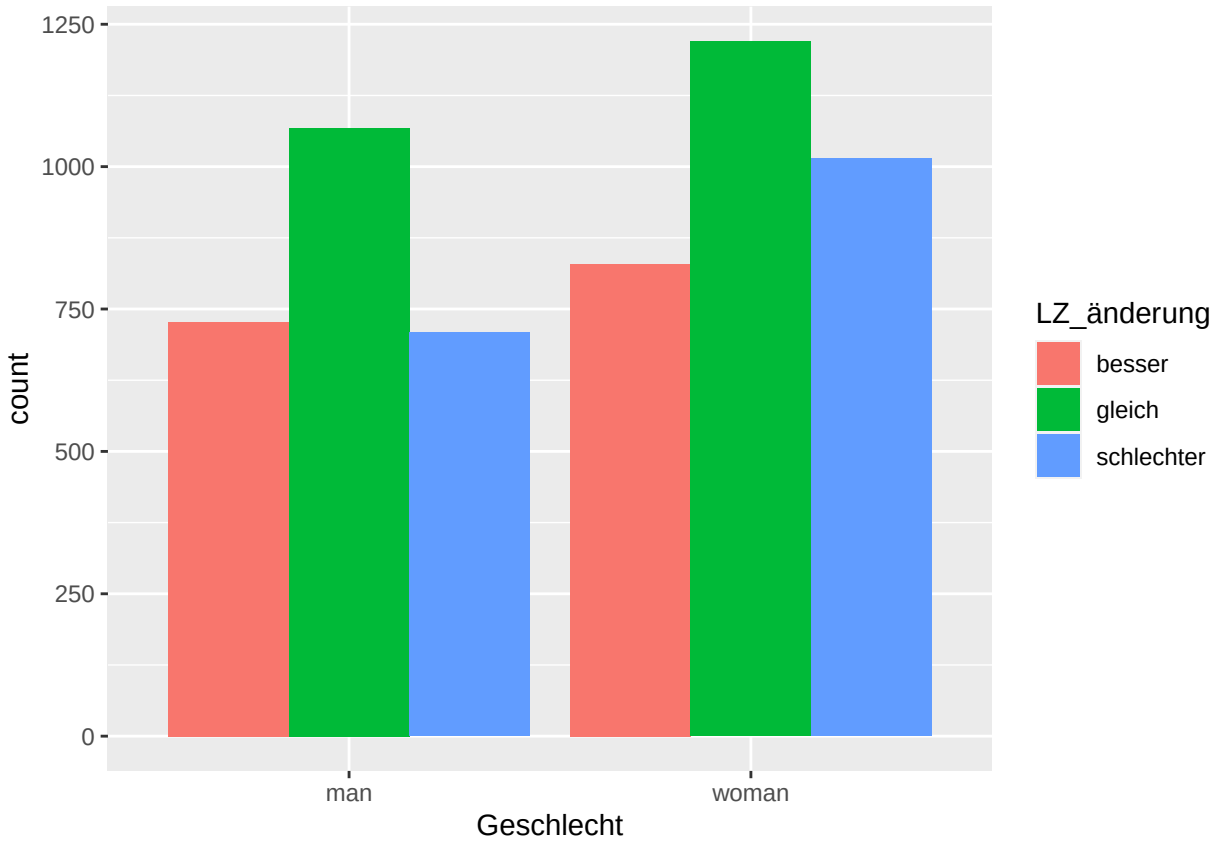

Figure 23: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf das Geschlecht

Diese Darstellung dient nicht der Beantwortung einer Hypothese und wurde aus reinem Interesse erstellt.

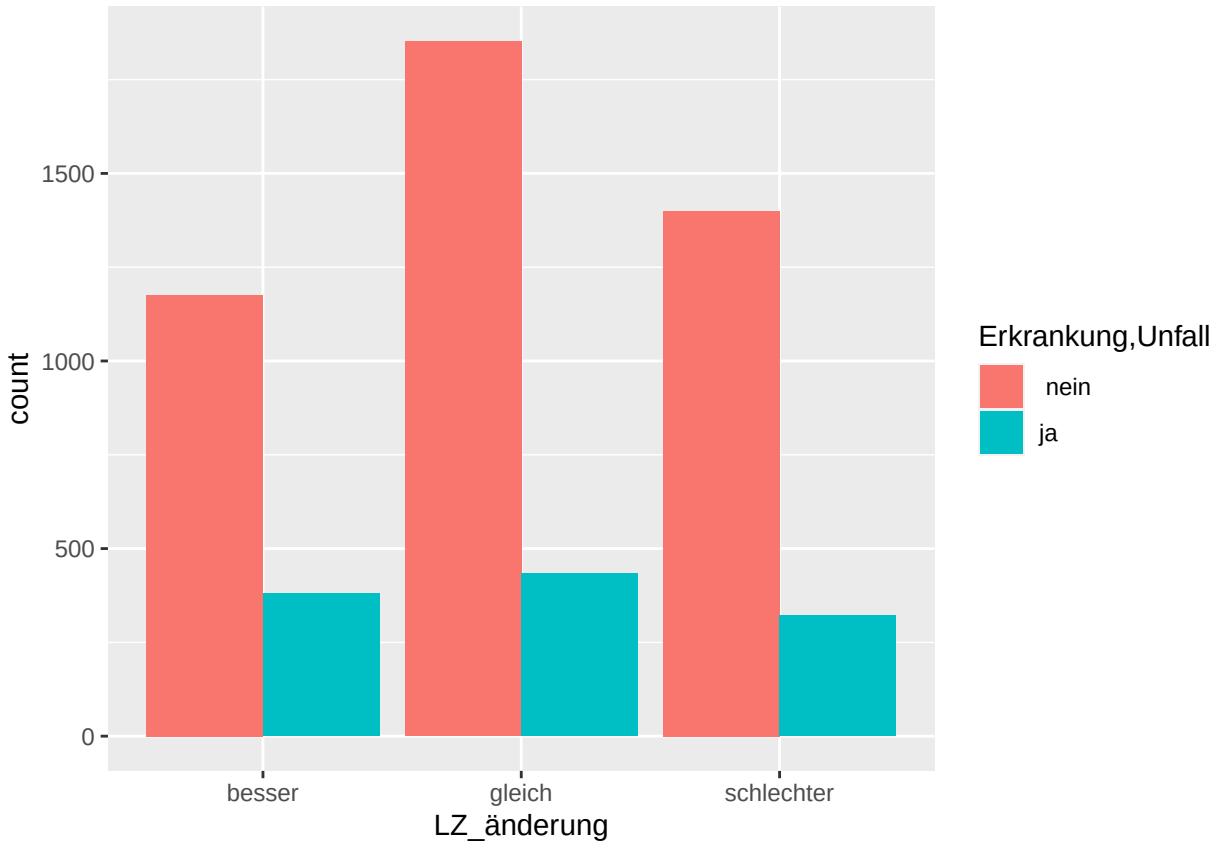

Figure 24: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf Erkrankung und Unfall

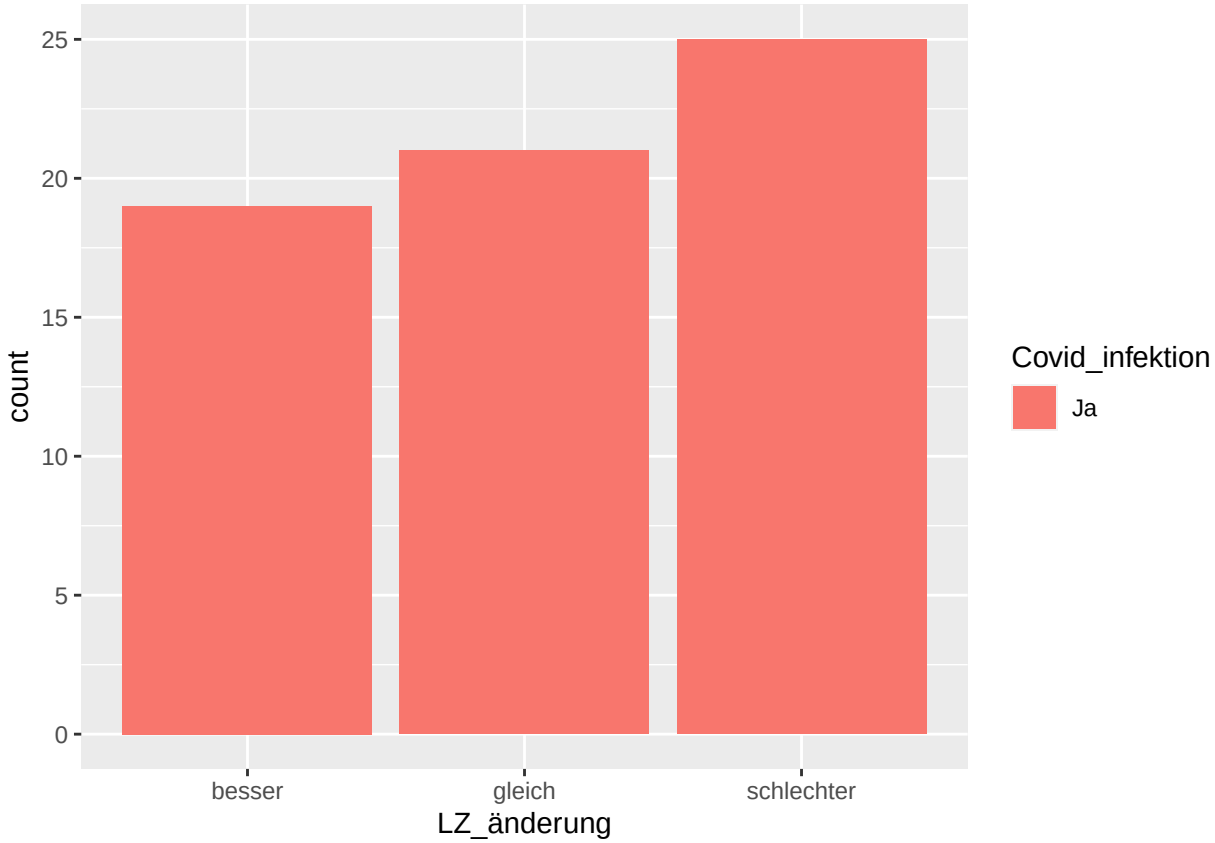

Figure 25: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf die Covid Infektion

Diese Grafiken zeigen die Veränderung der Lebenszufriedenheit im Bezug dazu, ob die befragte Person im letzten Jahr erkrankte oder einen Unfall erlebt hat, oder sich mit Covid infizierte. Die erste Grafik zeigt der Erste unserer eigenen Einflussfaktoren. Wir gingen davon aus, dass so ein Schicksal negative Folgen auf die Lebenszufriedenheit hat. Dies lässt sich anhand unserer Daten aber nicht verifizieren. Es gaben überraschenderweise mehr Leute, die diese Frage mit ja beantworteten, an eine Verbesserung der Lebenszufriedenheit erlebt zu haben. Der Unterschied ist jedoch vernachlässigbar klein. Die zweite Grafik verifiziert unsere Hypothese. Den Leuten, die sich mit Covid infiziert haben ging es tendenziell schlechter nach der Pandemie als vorher.

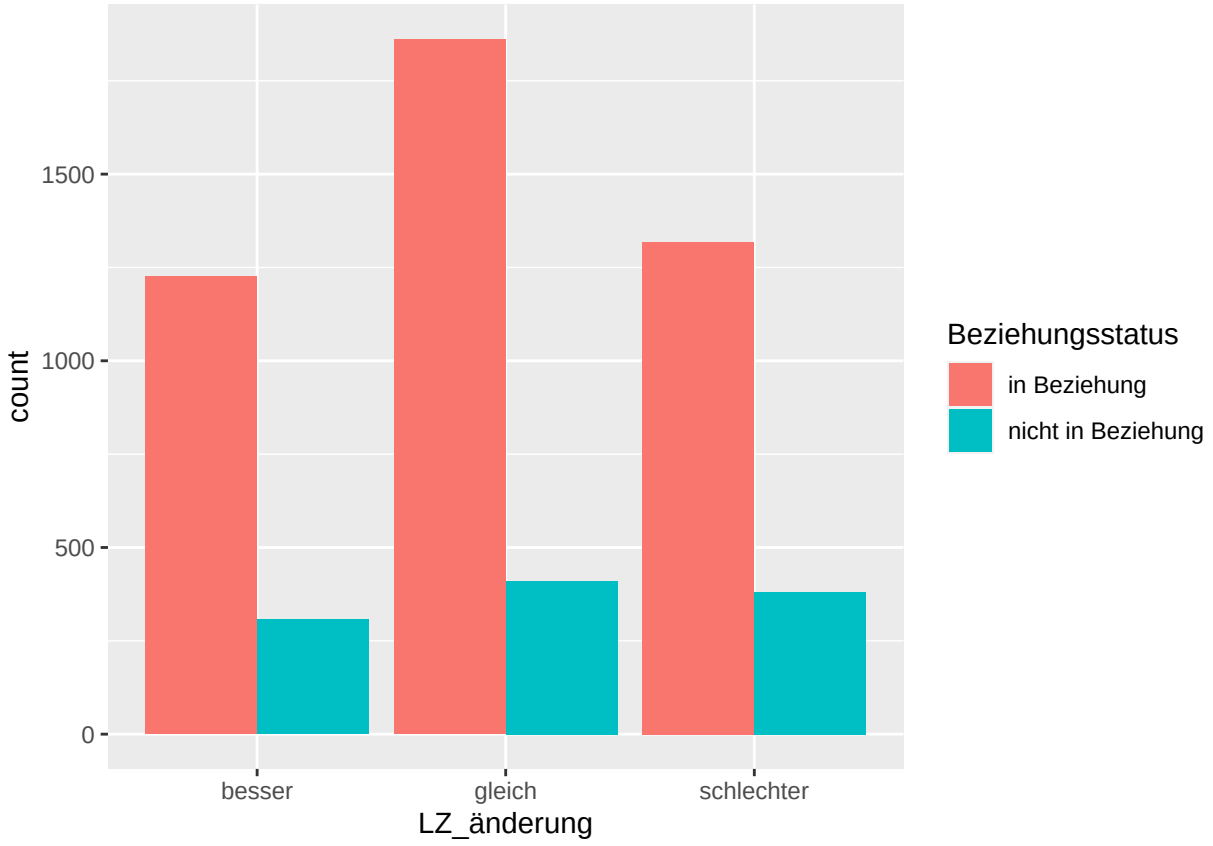

Figure 26: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf den Beziehungsstatus

Die Pandemie verbannte die Bevölkerung in ihre Häuser und erschwerte die Aufrechterhaltung eines sozialen Umfelds massiv. Für viele Menschen war Einsamkeit eine Folge davon. Wir dachten, dass besonders Menschen welche nicht in einer Beziehung waren davon betroffen wurden. Wie man der Grafik entnehmen kann, können wir das anhand unserer Daten aber nicht verifizieren.

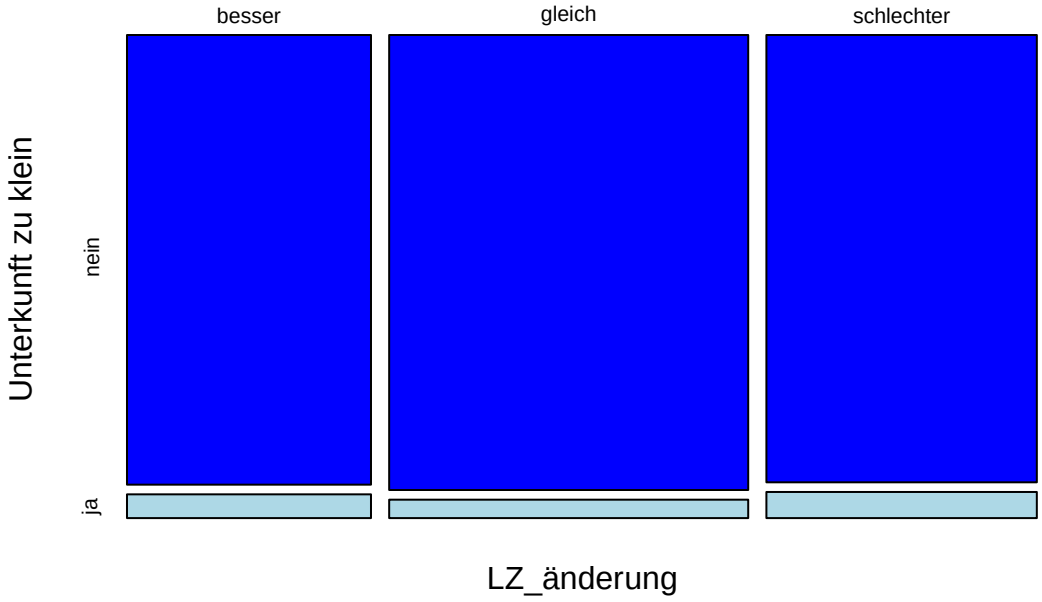

Figure 27: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf die Grösse der Unterkunft

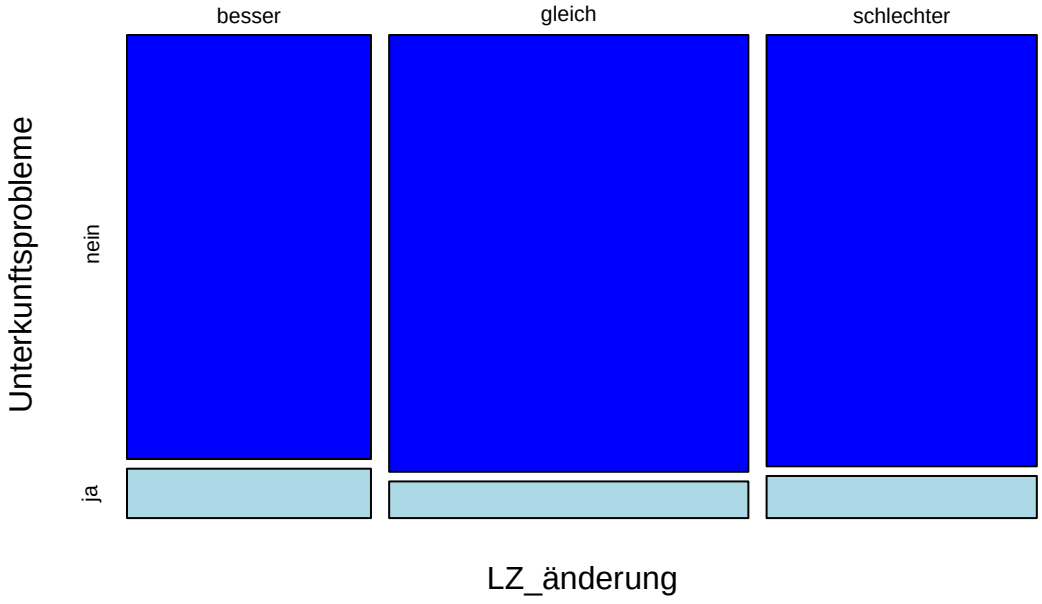

Figure 28: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf Probleme mit der Unterkunft

Diese beiden Mosaicplots zeigen die Auswertungen unserer anderer beiden eigenen Einflussfaktoren. Wir kamen aus eigener Erfahrung mit der Pandemie, und damit verbundenen Lockdowns, zum Schluss, dass Menschen die den Lockdown in einer Unterkunft verbringen mussten, welche entweder zu klein war oder andere Probleme aufwies, eine Abnahme der Lebenszufriedenheit rapportieren müssten. Diese Hypothese lässt sich nicht verifizieren, da die Daten wie man in der Grafik sehen kann keinen Unterschied zeigen zwischen Menschen welchen bei diesen Variablen ein “Ja” vermerkt haben.

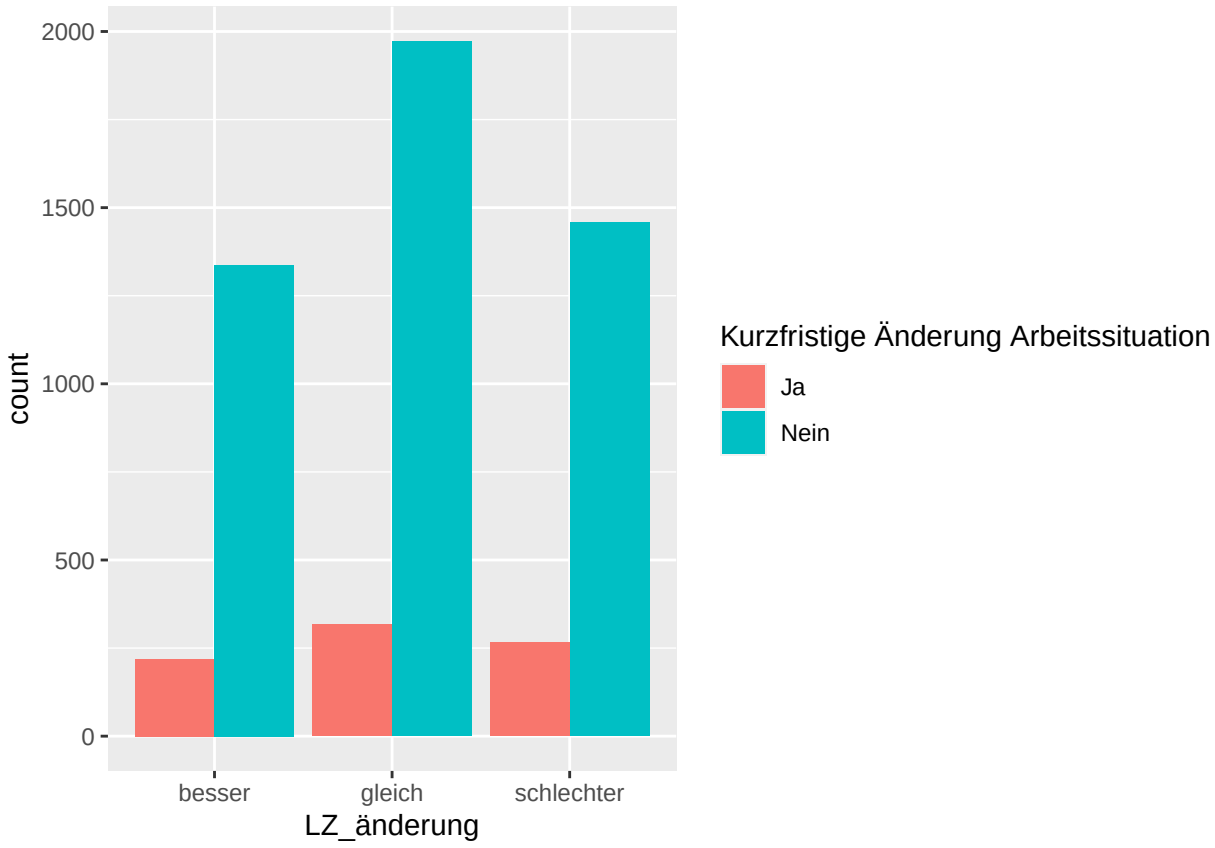

Figure 29: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf die kurzfristige Änderung der Lebenssituation

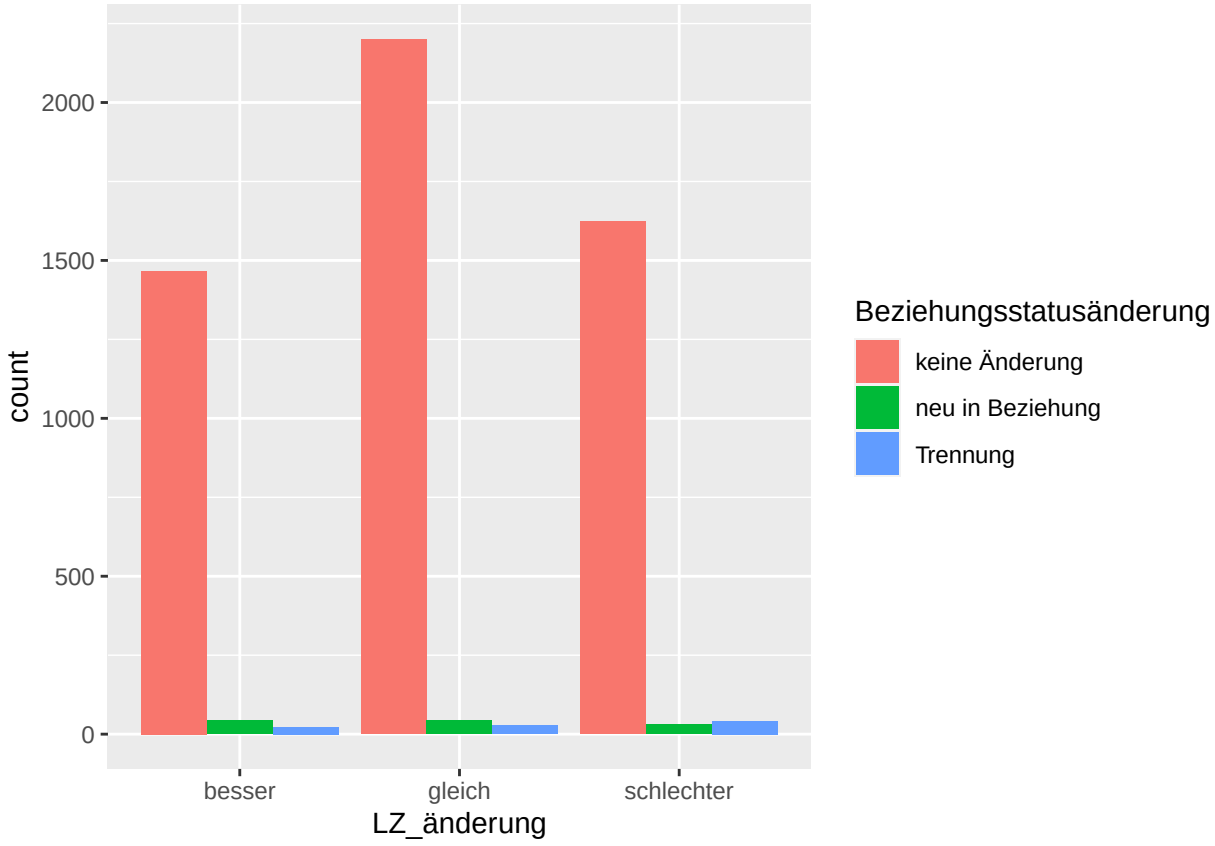

Figure 30: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf die Beziehungsstatusänderung

Unsere Hypothese lautete, dass Personen welche eine negative Veränderung in ihrem Leben erlebten während der Pandemie, negativ davon beeinflusst wurden. Diese zwei Grafiken stellen dar, wie sich die Lebenszufriedenheit von Personen, welche entweder eine Trennung oder eine Änderung der Arbeitssituation erlebten, veränderte. Unsere Hypothese lässt sich dadurch nicht verifizieren, da keine Unterschiede zwischen den Gruppen ersichtlich sind.

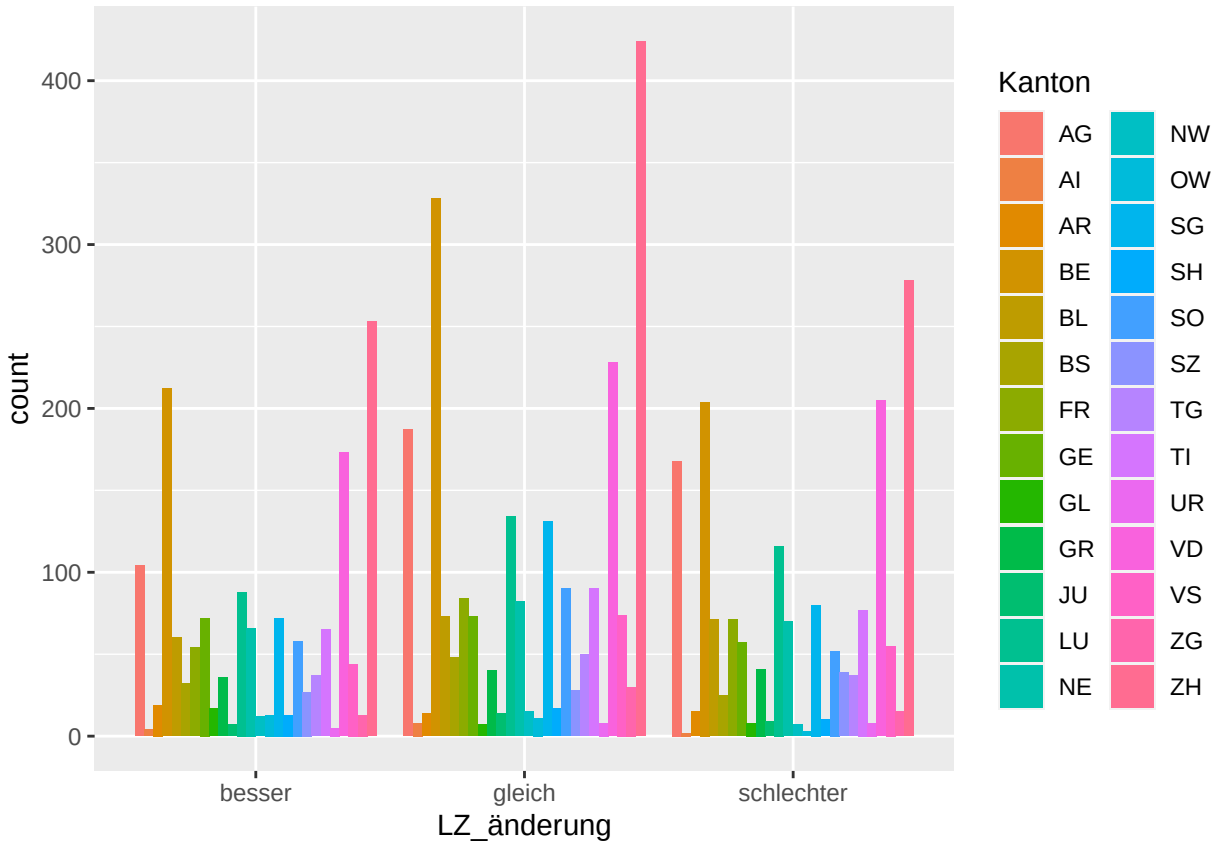

Figure 31: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf den Kanton

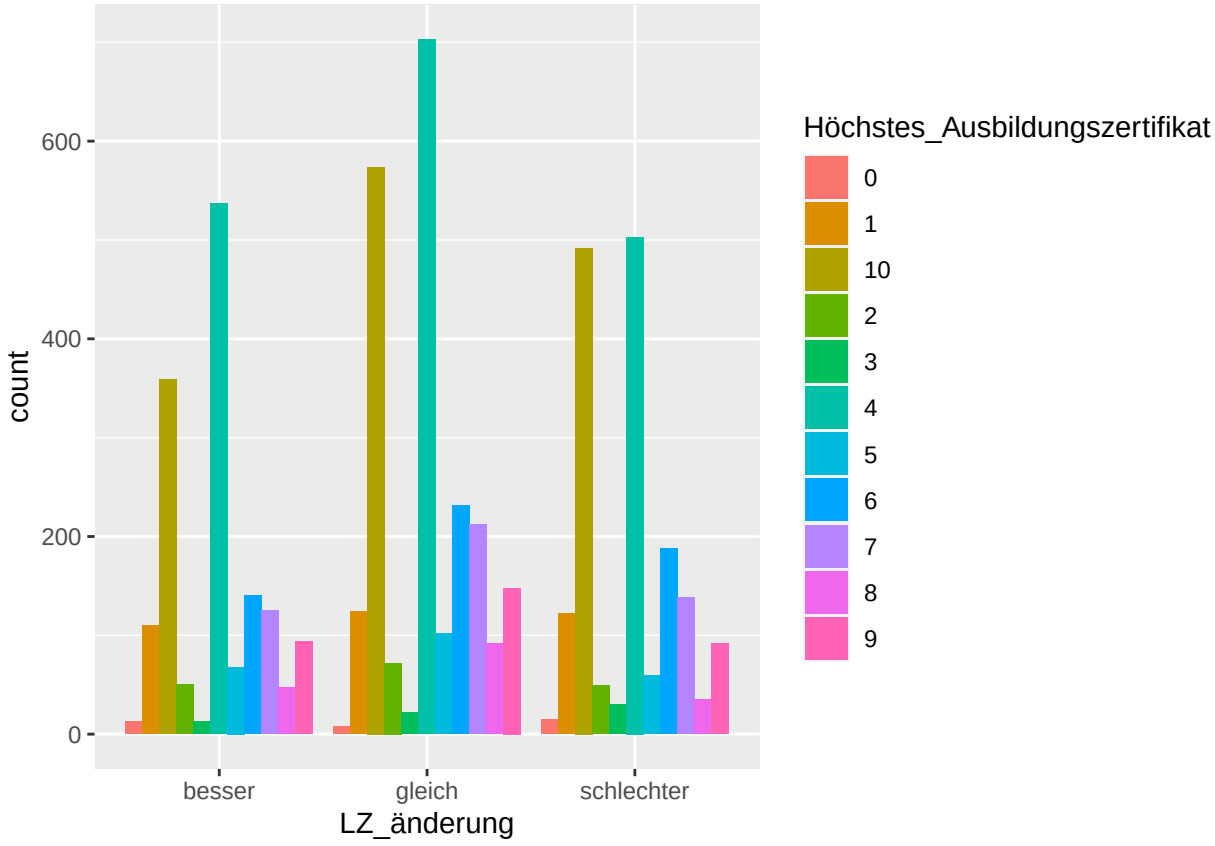

Figure 32: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf das Höchste Ausbildungszertifikat

Aus Platzgründen hier die Beschriftungen für die Werte in der Legende: 0 incomplete compulsory school 1 compulsory school, elementary vocational training 2 domestic science course, 1 year school of commerce 3 general training school 4 apprenticeship (CFC, EFZ) 5 full-time vocational school 6 bachelor/maturity 7 vocational high school with master certificate, federal certificate 8 technical or vocational school 9 vocational high school ETS, HTL etc. 10 university, academic high school, HEP, PH, HES, FH Diese beiden Grafiken dienen nicht zur Beantwortung einer Hypothese und wurden aus reinem Interesse generiert.

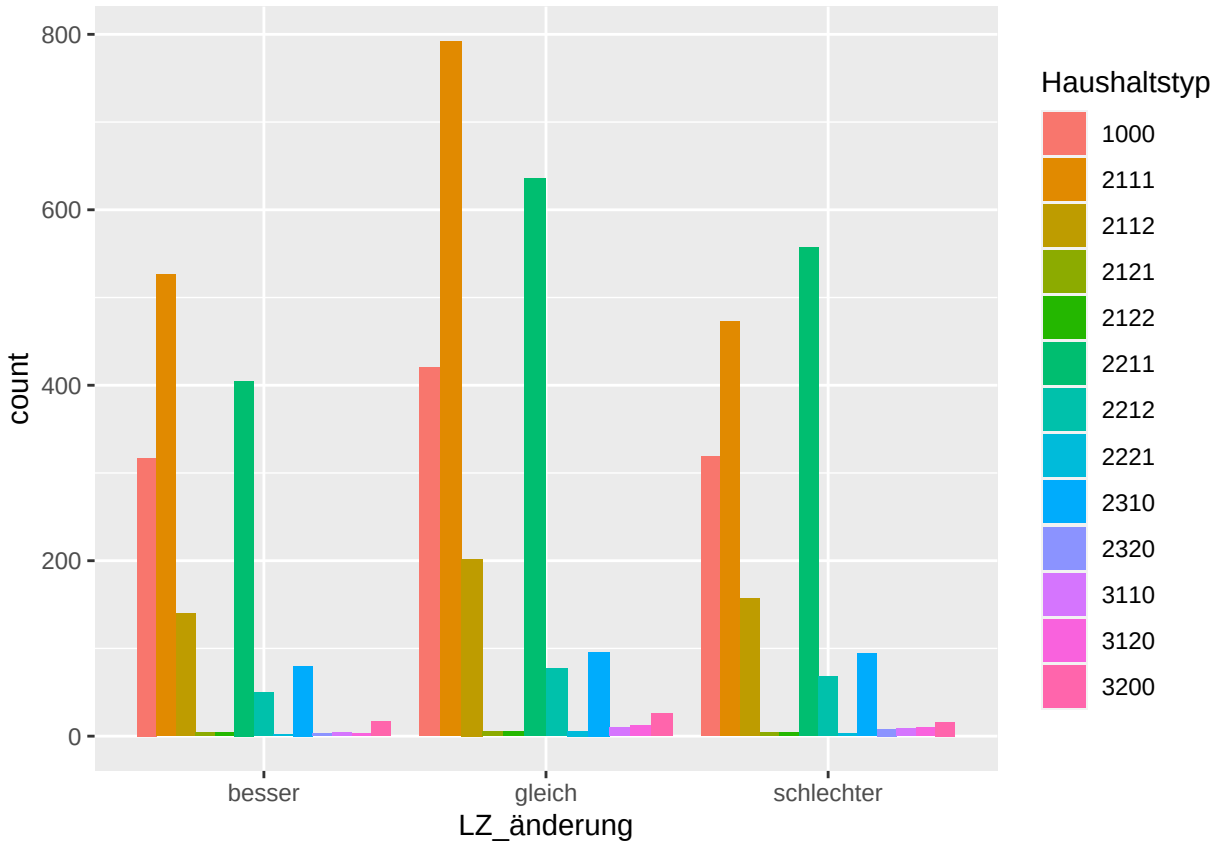

Figure 33: Dartellung zur Lebenszufriedenheitsänderung in Bezug auf den Haushaltstyp

Aus Platzgründen hier die Beschriftungen für die Werte in der Legende: 1000 One-person private households 2111 Married couple without children 2112 Consensual couple without children 2121 Married couple without children and another person 2122 Consensual couple without children and another person 2211 Married couple with children 2212 Consensual couple with children 2221 Married couple with children and another person 2222 Consensual couple with children and another person 2310 One parent with children 2320 One parent with children and another person 3110 Other types of households with only related family 3120 Other types of households with and without related family 3200 Other types of households without related family

Die letzte Fragestellung, welche wir verfolgten, war ob die Art des Haushaltes in denen sich die befragten Personen befanden einen Einfluss auf ihre Lebenszufriedenheit hatte. Wir formulierten dazu keine konkrete Hypothese und wollten die Grafik explorativ nutzen um mögliche Muster zu erkennen. Die Darstellung lässt aber keine Schlüsse ziehen.

Fazit

Die aufbereiteten Daten sind weder nach ihrer Grösse, noch der unter dem Kapitel “Plausibilität der Daten” erwähnten Punkte, repräsentativ. Sie spiegeln also die Bevölkerung nicht wieder und machen es demnach schwer bis unmöglich plausible und wissenschaftlich fundierte Schlüsse daraus zu ziehen. Ergo verlieren unsere bivariaten Darstellungen auch an Aussagekraft. Somit können wir keine unserer Hypothesen verifizieren oder falsifizieren. Die Resultate sind nicht eindeutig.