**Fundamentals of Data Visualization**

**Studentin: Julia Lobaton**

**Semester: 2. Semester**

**GitHub:**

**Dozentin:**

**LO1: Visualization basics, chart types**

**Einleitung**  
In dieser Lerneinheit geht es darum, die vielen verschiedenen Visualisierungstypen zu verstehen und zu lernen, wann welcher angewendet werden soll. Hierzu ist immer entscheidend, was mit der Visualisierung ausgedrückt werden möchte und wer das Zielpublikum ist.  
Die Abbildungen sind alle zusätzlich noch grösser dargestellt im Anhang zu finden.

**Daten**  
In diesem Kapitel werde ich die Daten von der Steinschlagrisiko-Challenge verwenden. Die Daten bestehen aus dem Datum, der Uhrzeit, der Masse und der Geschwindigkeit des Steinschlages. Diese Daten wurden an zwei Ablösungszonen erhoben. In der ersten Zone wurden 68 Steine und in der zweiten Zone 32 Steine registriert.

**Geschwindigkeit der Steine je nach Masse**Ziel von dieser Visualisierung war es, zu sehen wie die Geschwindigkeit in m/s je nach Masse der beiden Zonen verteilt ist. Ich wollte herausfinden, ob sich die beiden Zonen ähnlich verhalten oder nicht.  
Bei den Abbildungen 1 und 2 ist jeweils die Geschwindigkeit in Abhängigkeit der Masse pro Zone dargestellt. Damit die Visualisierungen gut verglichen werden können, habe ich die gleiche Skalierung gewählt.

*Abbildung 1: Geschwindigkeit in m/s der Steine je nach Masse für Zone 1 Abbildung 2: Geschwindigkeit in m/s der Steine je nach Masse für Zone 2*

**Auch habe ich in Abbildung 3 die beiden Zonen noch in einem Scatterplot dargestellt in zwei verschiedenen Farben.  
Bei der Ablösungszone 1 sind die Steine deutlich langsamer unterwegs, haben dafür zum Teil eine viel grössere Masse. Daraus könnte man die Annahme treffen, dass die Ablösungszone 1 weniger steil ist als die Ablösungszone 2 und somit die Steine öfters am Hang aufprallen und dadurch an Geschwindigkeit verlieren.   
Scatterplots sind sehr einfache Diagramme. Sie eignen sich besonders gut, wenn man die Beziehung zwischen zwei Variablen untersuchen möchte. Sie geben einem die Möglichkeit, die Korrelation visuell zu untersuchen. 1

*Abbildung 3: Geschwindigkeit in m/s je nach Masse für Zone kombiniert*

1 https://towardsdatascience.com/10-viz-every-ds-should-know-4e4118f26fc3

**Geschwindigkeit der Steine nach Zone**

Hier war mein Ziel, die Geschwindigkeit zu untersuchen, ohne Abhängigkeit der Masse der Steine. Hierfür habe ich einen Boxplot gewählt. Der Boxplot hat gegenüber dem in Abbildung 3 gezeigten Scatterplot, den Vorteil, dass aus einer Grafik sehr viele Informationen abgelesen werden können. Unter anderem den Median, das Minimum, das Maximum, die Spannweite und die Quantile.2 Das Minimum und das Maximum kann man ebenfalls auch beim Scatterplot einigermassen gut ablesen.  
Hier ist wie bei den Abbildungen 1-3 klar ersichtlich, dass die Steine aus der Ablösungszone 2 eine deutlich höhere Geschwindigkeit haben als die Steine aus der Ablösungszone 1. Wenn das Zielpublikum Menschen sind, welche sich mit dieser Materie nicht auskennen, finde ich eignet sich ein Scatterplot besser als ein Boxplot, da er meiner Meinung nach einfacher verständlich ist.

*Abbildung 4: Geschwindigkeit in m/s nach Ablösungszone*

**Tageszeitpunkt der Steinschläge**

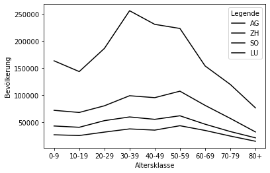
Nun wollte ich herausfinden, zu welchem Zeitpunkt die meisten Steine fallen. Den Tag habe ich folgendermassen aufgeteilt: Morgen (<= 06:00 bis < 12:00), Nachmittag (>=12:00 bis <18:00), Abend (>= 18:00 bis < 24:00) und Nacht (<=24:00 bis < 06:00).   
Der Zeitpunkt der Steinschläge kann übersichtlich mit einem Balkendiagramm visualisiert werden. Aus der Grafik kann man ablesen, dass die meisten Steine am Nachmittag fallen, gefolgt vom Abend. Am wenigsten Steine fallen in der Nacht.   
Da die Werte nahe beieinander liegen, eignet sich bei dieser Fragestellung ein Balkendiagramm sehr gut. Man kann auch kleine Unterschiede deutlich erkennen. Zum Vergleich habe ich bei Abbildung 6 noch ein Piechart mit denselben Daten visualisiert. Es ist klar zu sehen, dass das Balkendiagramm übersichtlicher ist und man die kleinen Unterschiede besser erkennen kann.3

*Abbildung 5: Tageszeit der Steinschläge Abbildung 6: Tageszeit der Steinschläge*

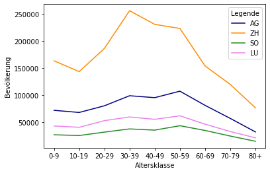
2 https://matheguru.com/stochastik/box-whiskers-plot.html 3 <https://www.data-to-viz.com/caveat/pie.html>

**LO2: Visual Perception**

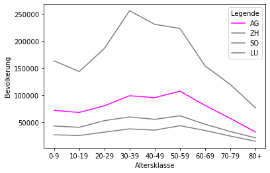
**Einleitung**  
In dieser Lerneinheit ist das Ziel, zu verstehen wie verschiedene Variablen die Wahrnehmung verändern können. Ich werde anhand von derselben Grafik zeigen, wie Farben, Orientierung, Grösse, Muster und Formen die Anschauung der Visualisierung verändern. Von Bertin’s Visual Variables habe ich mir fünf herausgesucht, welche ich genauer untersuchen möchte.4

**Daten**Ich werde in diesem Kapitel Daten von der «opendata.swiss» Seite benutzen. In dem Datensatz ist folgendes vorhanden: pro Kanton ist ersichtlich wie viele männliche und weibliche Menschen pro Altersklasse in diesem Kanton leben. Die Altersklassen sind folgendermassen unterteilt: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+. Ich habe den Datensatz gefiltert und werde nur die Kantone Luzern, Aargau, Zürich und Solothurn betrachten.  
Dies ist meine Grafik, bei welcher man noch keine Unterschiede oder Erkenntnisse machen kann. Diese werde ich je nach Variabel verändern und anpassen.

*Abbildung 7: Grafik vor Anpassung*

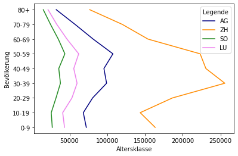
**Farben**In der Datenvisualisierung gibt es folgende Anwendungsfälle für Farben: Datengruppen können voneinander unterschieden werden, Datenwerte können visualisiert werden und mit Farben können einzelne Datenwerte hervorgehoben werden. Oft werden Farben benutzt um einzelne Elemente, welche keine Ordnung haben, in der Visualisierung zu unterscheiden. Hier ist es wichtig, Farben zu wählen, welche deutlich voneinander zu unterscheiden sind und keine der Farben im Vergleich zur anderen hervorsticht. Auch ist es wichtig, dass die Farbe keinen Eindruck von Ordnung geben.5 Auch muss man darauf achten, dass die Grafik für Farbenblinde problemlos zu lesen ist. Dies habe ich ebenfalls bei der Farbwahl berücksichtigt.6 Aufgrund von diesen Erkenntnissen habe ich meine Grafik angepasst. In der Abbildung 8 sind die Farben so gewählt, dass keine hervorsticht und alle gut voneinander zu unterscheiden sind. 7

*Abbildung 8: Grafik mit äquivalenten Farben*

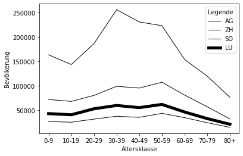
**In der Abbildung 9 habe ich die Farben so gewählt, dass es die Linie vom Kanton Aargau hervorhebt.

*Abbildung 9: Grafik mit Farben, dass Aargau hervorgehoben wird*

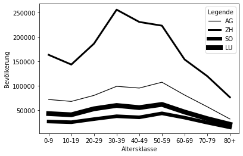
4 <https://www.e-education.psu.edu/geog486/node/637>  
5 <https://clauswilke.com/dataviz/color-basics.html>  
6 https://www.extensis.com/de-de/blog/so-gestalten-sie-designs-f%C3%BCr-farbenblinde  
7 https://www.acrylfarben.ch/farbenlehre-der-farbkreis-nach-johannes-itten/

**Orientierung**Bei der Abbildung 10 wurden die Achsen vertauscht. Die Grafik ist für den Leser schwieriger zu interpretieren, da man es sich gewohnt ist, dass auf der x-Achse die unabhängige Variabel stehen.8

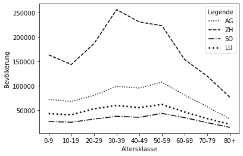
*Abbildung 10: x- und y-Achsen vertauscht*

**Grösse**Bei der Abbildung 11 wollte ich den Fokus auf den Kanton Luzern setzen. Dafür eignet sich die Anpassung der Grösse sehr gut. Dafür habe ich die Linie von Luzern dicker dargestellt als die anderen drei Linien.

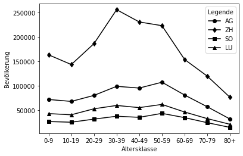
*Abbildung 11: Grösse einer Linie verändert*

Bei der Abbildung 12 haben alle vier Linien eine andere Dicke. Die Kantone nur anhand der Liniendicke zu unterscheiden, eignet sich in diesem Beispiel nicht sehr gut. Bei der Abbildung 12 ist bei den unteren beiden Linien nicht auf den ersten Blick erkennbar, welche Linie zu welchem Kanton gehört.

*Abbildung 12: Grösse aller Linie verändert*

**Muster**Die unterschiedlichen Linientypen können ebenfalls dazu benützt werden, die vier Kantone zu unterscheiden. Dies hat den grossen Vorteil gegenüber der Unterscheidung durch Farben, dass wenn man die Visualisierung ausdruckt, die einzelnen Linien trotzdem unterscheiden und zugeordnet werden können.

*Abbildung 13: Linienstyle geändert*

**Formen**Mit Formen werden die einzelnen Datenpunkte dargestellt.9 In diesem Beispiel (bei Abbildung 14) ist dies jedoch nicht wichtig, da die Kategorien des Alters nicht weiter unterteilt sind.

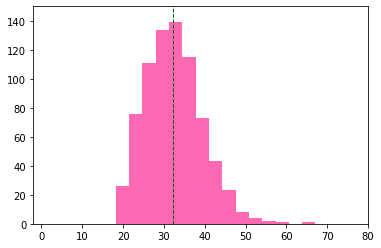
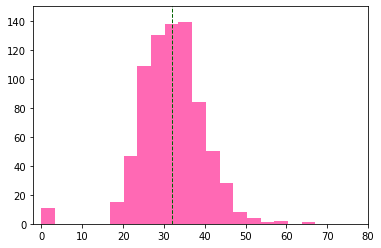
*Abbildung 14: Formen bei Datenpunkten*

8https://de.wikipedia.org/wiki/Abh%C3%A4ngige\_und\_unabh%C3%A4ngige\_Variable#:~:text=Im%20Falle%20einer%20einzelnen%20Variablen,und%20x%20die%20unabh%C3%A4ngige%20Variable9 https://gertingold.github.io/pythonnawi/graphics.html

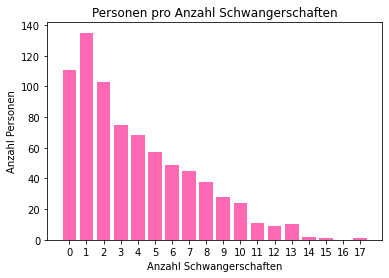
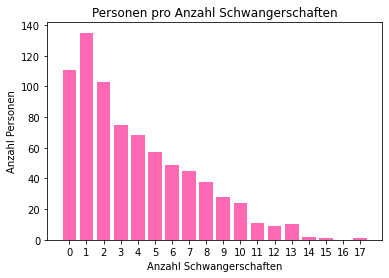
**LO3: Design Principles vs. Data**

**Einleitung**In diesem Kapitel werde ich aufzeigen, wie die Datenvorverarbeitung in Verbindung mit den Visualisierungen steht. Die meisten Daten müssen noch bearbeitet oder verändert werden, bevor man damit aussagekräftige Visualisierungen machen kann. Während dieser Vorbereitung werden wichtige Entscheidungen getroffen, welche die Visualisierung beeinflussen.   
Es ist wichtig, dass man weiss, mit welchen Datentypen man es zu tun hat. Daten können in qualitative oder quantitative Merkmale eingeteilt werden. Diese beiden Kategorien haben noch zwei Untergruppen. Qualitative Daten können in nominal oder ordinal eingeteilt werden. Quantitative Daten kann man in diskrete und stetige Merkmale einteilen.10Es ist ausserdem wichtig, dass man die Daten zuverlässig untersucht und bereinigt. Dazu gehört zum Beispiel die Untersuchung der Ausreisser. Man muss abschätzen, ob es sich bei diesen Zahlen um Fehler handelt, oder ob es sich tatsächlich um Ausreisser handelt. Auch kann man irrelevante Daten, welche für die Untersuchung nicht wichtig sind, löschen. Ein weiterer Punkt, den man beachten muss, ist, dass man kontrolliert, dass jede Spalte den richtigen Datentyp hat. Konkret heisst dies, zu schauen, ob zum Beispiel ein Daum auch als solches gespeichert ist oder eine Zahl als integer oder float.11

**Daten**In dieser Lerneinheit werde ich mit einem Datensatz aus einer Minichallenge im Modul Data Wrangling arbeiten. Diese Datei enthält Daten von 767 Frauen, welche mindestens 21 Jahre alt sind. Diese wurden auf verschiedene Merkmale getestet und es wurde untersucht, ob sie an Diabetes leiden oder nicht. Es sind unter anderem Anzahl Schwangerschaften, Blutdruck, Insulin, BMI, Alter und noch einige weitere Resultate von Untersuchungen hinterlegt. Der Datensatz enthält auch die Information, ob die Person Diabetes hat oder nicht. Der Datensatz enthält quantitative diskrete Variablen zum Beispiel Anzahl Schwangerschaften. Die anderen Daten wie zum Beispiel: Alter, Blutdruck und BMI sind quantitativ stetige Variablen.

**Darstellung quantitativ stetige Datentypen**Im Datensatz hat es Werte, welche keinen Sinn ergeben. Zum Beispiel hat es beim BMI zum Teil den Wert 0. Dies kann nicht sein. Beim BMI handelt es sich um einen quantitativen stetigen Datentyp. Histogramme sind eine Standartmethode, um quantitativ stetige Variablen zu visualisieren. Ein Histogramm zeigt die Verteilung der Werte.12 Bei Abbildung 15 sieht man, wie das Histogramm mit den Nullwerten aussieht. Die Nullwerte habe ich durch NaN-Werte ersetzt, so werden sie für das Histogramm nicht berücksichtigt. Bei Abbildung 16 ist das Histogramm mit NaN-Werten. Die grüne Linie ist bei beiden Abbildungen der Median. Dieser verändert sich kaum. Bei Abbildung 15 ist er 32 und bei der Abbildung 16 ist er 32,3. Daraus lässt sich schliessen, dass es nur wenige Null Werte hat.

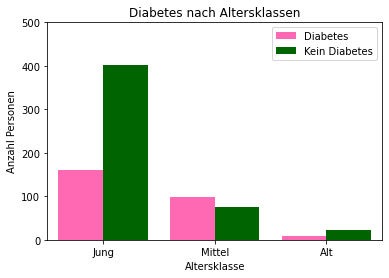
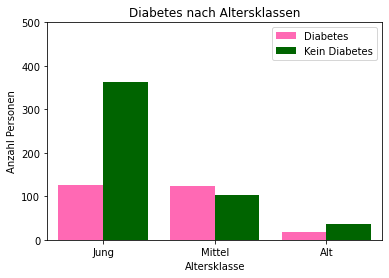
*Abbildung 15: BMI Werte mit 0 Werten Abbildung 16: BMI Werte mit NaN-Werten*

10 <https://blog.minitab.com/de/qualitative-quantitative-attributive-diskrete-und-stetige-daten>  
11 <https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4>  
12 https://statisticsbyjim.com/basics/data-types/  
**Darstellung quantitativer stetiger Variablen**  
Quantitative diskrete Daten sind eine Zählung eines vorhandenen Merkmals, einer Ereignisses oder eines Elements. Es gibt nur eine endliche Anzahl an möglichen Werten. Um diskrete Variablen darzustellen, eignet sich ein Balkendiagramm. Jeder der Balken steht für einen Wert und die Höhe stellt den Anteil an den gesamten Daten dar 13. Mit dem vorhandenen Datensatz wollte ich ****herausfinden, wie viele Personen es pro Anzahl Schwangerschaften hat.

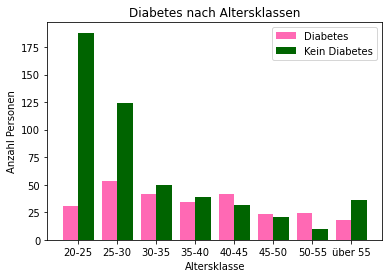
*Abbildung 17: Personen pro Anzahl Schwangerschaften*

**Diabetes verteilt auf die Altersklassen**Ich wollte herausfinden, wie das Alter im Zusammenhang mit Diabetes steht. Dafür habe ich das Alter in drei Kategorie unterteilt. Um zu zeigen, dass die Vorbereitung der Daten einen Einfluss auf die Visualisierung hat, habe ich die Unterteilung der Altersklassen von Abbildung 18 zu Abbildung 19 verändert.

Einteilung Altersklassen:  
Abbildung 17: Jung = von 20 bis 34,9 / Mittel = von 35 bis 54,9 / Alt = ab 55  
Abbildung 18: Jung = von 20 bis 39,9 / Mittel = von 40 bis 59,9 / Alt = ab 60

****

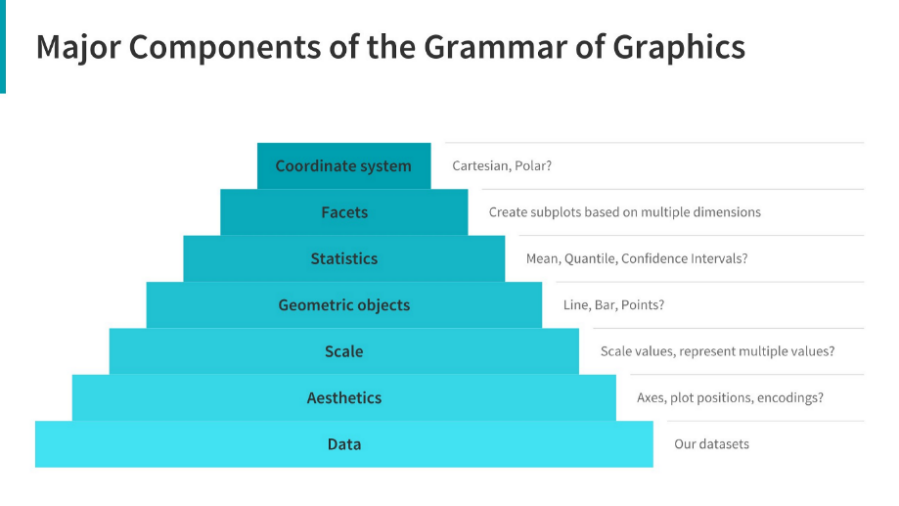
*Abbildung 18: Erste Variante der Altersgruppen Abbildung 19: Zweite Variante der Altersgruppen*

****Wenn man die Abbildungen 18 und 19 vergleicht, wird einem schnell klar, dass die Datenvorbereitung eine wichtige Rolle spielt und einen Einfluss auf die Visualisierung hat.  
Somit hat man je nach Vorbereitung der Daten ein anderes Bild und interpretiert die Daten anders.  
In Abbildung 20 habe ich acht Altersgruppen gebildet. Je nach dem, was das Ziel der Visualisierung ist, muss man die Datenvorbereitung anpassen.

*Abbildung 20: Acht Altersgruppen*

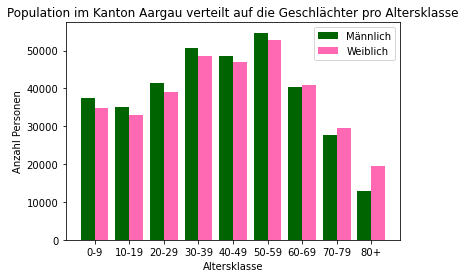
13 https://statisticsbyjim.com/basics/data-types/

**LO 4: Grammar of graphics tools**

**Einleitung**In dieser Lerneinheit werde ich aufzeigen, wie «The Grammar of Graphics» die Erstellung einer Visualisierung beeinflusst. Einzelne Teilgebiete wurden in andere Lerneinheiten bereits aufgezeigt. In diesem Kapitel werde ich eine Übersicht über das komplette Thema geben.  
Ich werde mich am Aufbau wie in der Abbildung 21 orientieren.13

*Abbildung 21: Major Components of the Grammar of Graphics*

**Data**Hier gibt es eine erste Überschneidung mit dem LO 3. Ich werde mit den «opendata.swiss» Daten aus LO 2 arbeiten.

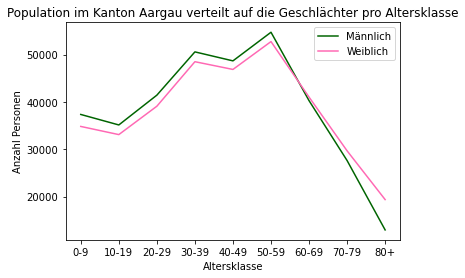
**Aesthetics**Hier geht es darum, zu entscheiden welche Attribute die Achsen annehmen sollen und wie diese dargestellt werden sollen.Auch stellt sich die Frage, wie viele Dimensionen der Plot haben soll. Ein Plot mit zwei Dimensionen wäre zum Beispiel ein schwarz-weisser Scatter Plot. Als dritte Dimension könnte man zum Beispiel mit Farben arbeiten und dann als vierte noch die Grössen der Punkte verändern. 13

*Abbildung 22: Wie viele Männer und Frauen hat es pro Altersklasse (Barplot)*

Da ich aber herausfinden wollte, wie viele Männer und Frauen es pro Altersklasse hat, habe ich eine weitere Dimension dazu genommen, die Farbe.   
Eine Möglichkeit meine Fragestellung zu beantworten, ist mit einem Balkendiagramm. An der Abbildung 22 ist zu erkennen, dass es bei den jüngeren Menschen mehr Männer hat und bei den älteren Menschen mehr Frauen.

13 https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149

**Scale**Mit der Skalierung entscheidet man, wie man die X- und die Y- Achse skalieren möchte. Die Skalierung kann beliebig nach Bedürfnissen und Fragestellung angepasst werden.13

**Geometric objects**Hier wird entschieden, mit welcher Art Grafik die Information dargestellt wird. Zum Beispiel mit Punkten, mit Linien oder mit einem Barplot.13 Die gleiche Fragestellung, welche mit Abbildung 22 beantwortet wurde, kann auch mit einem Liniendiagramm beantwortet werden (Abbildung 23).

*Abbildung 23: Wie viele Männer und Frauen hat es pro Altersklasse (Liniendiagramm)*

**Statistics**Je nach Fragestellung kann es von Bedeutung und Nutzen sein, wenn man statistische Werte wie zum Beispiel den Median oder den Durchschnitt in der Grafik markiert.13 Bei den Abbildungen 22 und 23 macht hätte man davon keinen grossen Nutzen. Bei Abbildung 15 und 16 habe ich mit dem Median gearbeitet und diesen in beiden Grafiken eingezeichnet.

**Facets**Wenn viele Variablen miteinander dargestellt werden wollen, kann eine Grafik schnell unübersichtlich werden. In solchen Fällen lohnt es sich, die den Plot in Facets aufzuteilen. So kann der Plot übersichtlich aufgeteilt und dargestellt werden.14

**Coordinate system**Es stellt sich die Fragen, welches Koordinatensystem man verwenden möchte. Soll es kartesisch oder polar sein? Wenn man zum Beispiel viele Kategorien in einem Bar Chart hat, kann es übersichtlicher werden, wenn man dies in einem Polarkoordinatensystem darstellt.15

14 https://bookdown.org/joone/ComputationalMethods/daten-visualisieren.html

15 https://medium.com/optima-blog/using-polar-coordinates-for-better-visualization-1d337b6c9dec

**Anhang**

**LO1**

Abbildung 1: Geschwindigkeit in m/s der Steine je nach Masse für Zone 1

****

Abbildung 2: Geschwindigkeit in m/s der Steine je nach Masse für Zone 2



Abbildung 3: Geschwindigkeit in m/s je nach Masse für Zone kombiniert

Abbildung 4: Geschwindigkeit in m/s nach Ablösungszone

Abbildung 5: Tageszeit der Steinschläge

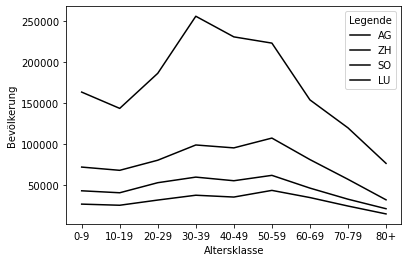


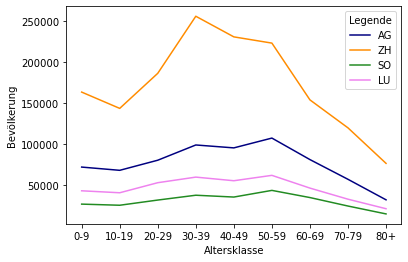
Abbildung 6: Tageszeit der Steinschläge

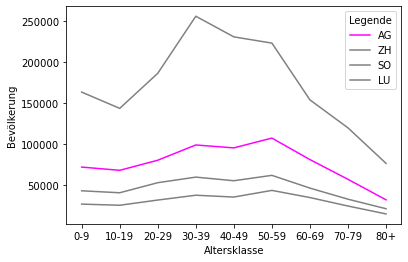


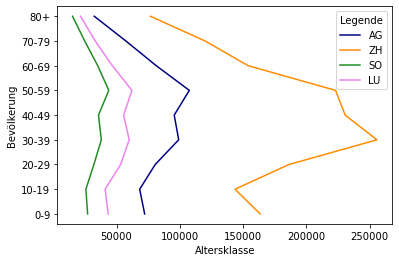
**LO 2**

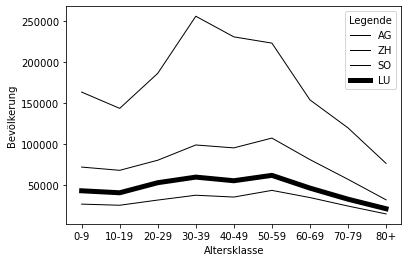
Abbildung 7: Grafik vor Anpassung

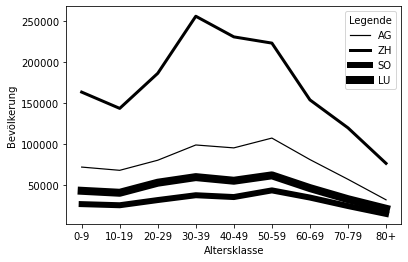


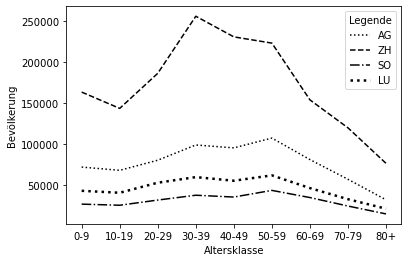
Abbildung 8: Grafik mit äquivalenten Farben

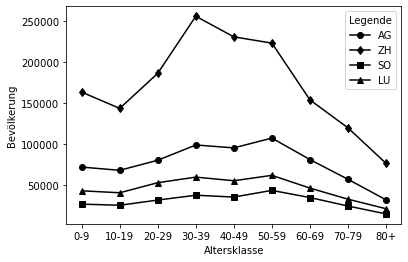
Abbildung 9: Grafik mit Farben, dass Aargau hervorgehoben wird

Abbildung 10: X- und Y-Achsen vertauscht

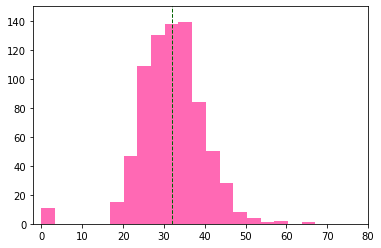
Abbildung 11: Grösse einer Linie verändert

Abbildung 12: Grösse aller Linien verändert

Abbildung 13: Linienstyle geändert

 Abbildung 14: Formen bei Datenpunkten

**LO 3**

Abbildung 15: BMI Werte mit 0 Werten

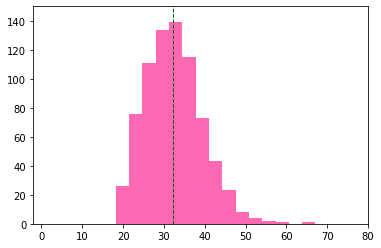
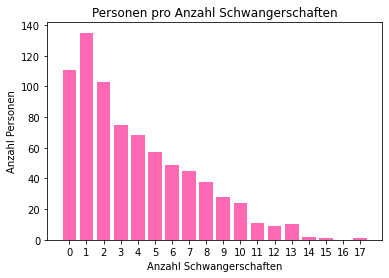
Abbildung 16: BMI Werte mit NaN-Werte

Abbildung 17: Personen pro Anzahl Schwangerschaften

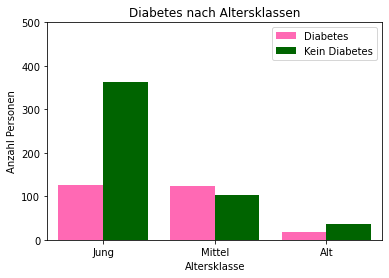
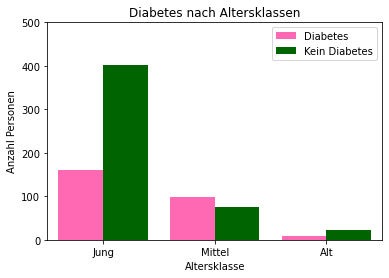
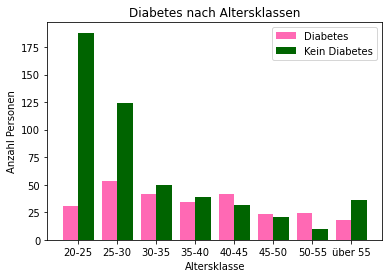
Abbildung 18: Erste Variante der Altersgruppen

Abbildung 19: Zweite Variante der Altersgruppen



Abbildung 20: Acht Altersgruppen