Grundlagen Datenvisualisierungen

Bericht

Student: Florin Barbisch

Studiengang: FHNW, Data Science

Fachexpertin: Arzu Cöltekin

Datum: 4.6.2022

Table of Contents

[LE1 3](#_Toc97822980)

[Balkendiagramm 3](#_Toc97822981)

[Liniendiagramm 3](#_Toc97822982)

[Histogramm 4](#_Toc97822983)

[Punktewolken/Heatmap 4](#_Toc97822984)

# LE1

Als Datensatz verwende ich die Wetterdaten der Wetterstation Mythenquai der Seepolizei Zürich (Wasserschutzpolizei et al., 2021) aus der Wettermonitor-Challenge (Brönnimann, 2021), welche ich letztes Jahr absolviert habe. Der Datensatz erhält Messwerte von Wetterdaten über mehrere Jahre in 10-Minuten Abständen.

Aufgabe in der Challenge war es, ein Dashboard für die aktuellen Daten zu erstellen. Da es für diesen Use-Case nicht viel Plots gibt, möchte ich als Use-Case eine explorative Datenanalyse durchführen. Die Ergebnisse sollten aber trotzdem von einem Segler verstanden werden.

Einfachheitshalber analysiere ich nur die Daten von einem Jahr, da ich mich hauptsächlich mit den verschiedenen Diagrammtypen auseinandersetzen möchte.

Alle Diagramme sind meine Eigenen. Der Quellcode für diese Diagramme befindet sich auf einem GitHub-Repository (Barbisch, 2022).

## Balkendiagramm

Chart, bar chart

Description automatically generatedBei einem Balkendiagramm lässt sich eine numerische Variable über verschiedene Kategorien vergleichen. Die Balken können entweder vertikal oder horizontal sein. Als Kategorien kommen auch Intervalle von stetigen Zeitvariablen in Frage (Monat, Quartal, Jahr…). Bei Intervallen von anderen stetigen Variablen eignet sich eher ein Histogramm (z.B. Körpergrösse oder IQ). Bei solchen Intervallen (und ordinalen Kategorien) ist wichtig, dass die Balken nach der stetigen Variable sortiert sind. Wenn man nominale Kategorien hat, macht eine Sortierung nach der Balkengrösse Sinn. In der Abbildung 1 ist ein Balkendiagramm zu sehen. Ein Balken repräsentiert jeweils ein Zeitintervall von einem Monat. Die Höhe der Balken gibt den Niederschlag in mm an, dies kann an der y-Achse abgelesen werden. Die Balken sind hier nach Monat sortiert, so lässt sich der Verlauf über das ganze Jahr leicht erkennen.

Abbildung 1: Balkendiagramm über den Niederschlag in Mythenquai pro Monat im Jahr 2018

## Liniendiagramm

Chart, line chart

Description automatically generatedEin Liniendiagramme hat Ähnlichkeiten mit dem Balkendiagram. Da aber statt Balken eine Linie gezeichnet wird, setzt dies einen Zusammenhang der Punkte in x-Richtung voraus (meistens ein temporaler). Dies macht es im Vergleich zu einem Balkendiagramm leichter Trends zu erkennen (Chynał & Sobecki, 2016, p. 163) und des Weiteren macht es keinen Sinn kategorielle Daten auf der x-Achse darzustellen.

Da die Werte der x-Achse einen Zusammenhang mit dem Wert links und rechts davon haben und die Daten in regelmässigen Abständen/Intervallen erhoben wurden, muss nicht für jeden Datenpunkt eine neue Achsenbeschriftung erstellt werden. Die x-Werte zwischen den Intervallbeschriftungen lassen sich dann herleiten.

Abbildung 2: Liniendiagramm über den Niederschlag in Mythenquai pro Monat im Jahr 2018

In der Abbildung 2 wird wieder der Niederschlag in Mythenquai pro Monat im Jahr 2018 dargestellt. Statt Balken werden Linien zwischen den einzelnen Datenpunkte gezogen. Dies erschwert das Ablesen von quantitativen Werten (Steedle, 2010). Besonders im Juli und August ist es schwer zu erkenne wo sich der Datenpunkt befindet. Deshalb ist für diese Visualisierung das Balkendiagramm besser geeignet.

## Histogramm

In einem Histogramm kann eine Verteilung einer kontinuierlichen Variable dargestellt werden. Es werden ähnlich wie im Balkendiagramm Balken gezeichnet, die Höhe der Balken entspricht aber der Anzahl Werte, welche im Intervall des Balken vorkommen. Die x-Achse muss sortiert sein, sonst lässt sich keine Verteilung erkennen.

Beim Histogramm muss man Klassengrössen wählen. Die Klassengrösse entspricht dem Intervall eines Balken. Je grösser die Klassengrösse, je weniger Balken gibt es und mehr Informationen werden versteckt. Wenn man zu wenig Balken hat, kann man die Art der Verteilung (z.B. normalverteilt) nicht erkennen. Wenn man zu viele Balken hat, wird das Rauschen der Daten sichtbar. Hier gibt es nicht eine Formel welche immer funktioniert. (Sturges, 1926) hat aber einen systematischen Ansatz mit der Formel entwickelt um die Anzahl Balken zu ermitteln.

Abbildung 3: Histogramm über die Verteilung des totalen Niederschlags pro Tag in Mythenquai

In der Abbildung 3 ist eine Verteilung des Niederschlages in Mythenquai zu sehen. Der Betrachter sieht hier direkt, dass es an vielen Tagen nicht bis wenig regnet und nur an ganz wenigen Tagen viel regnet.

# Boxplot

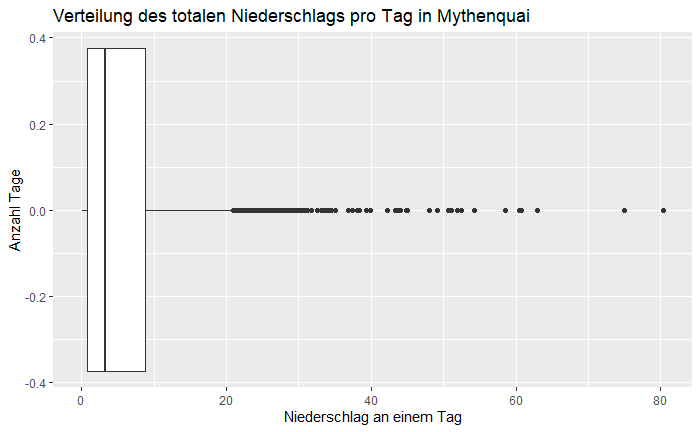
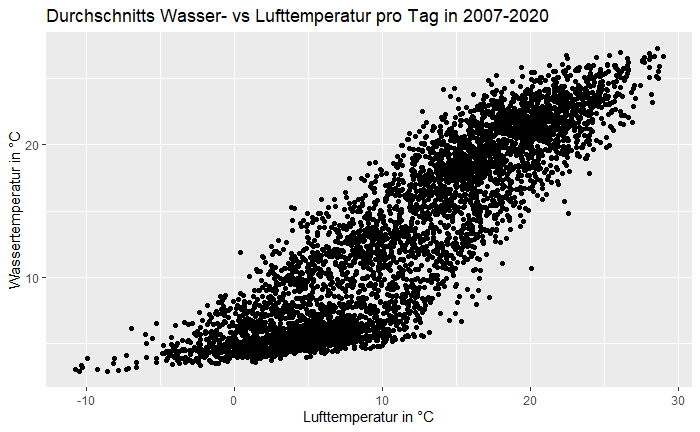
Eine Alternative zum Histogramm ist der Boxplot. Dort werden in einer mit einer Box die mittleren 50% der Daten markiert. Der Strich in der Box zeigt den median an. Die Antennen in beide Richtungen sind 1,5-mal so lang wie der Interquartilsabstand. Punkte welche ausserhalb dieser Box und der Antenne liegen werden einzeln markiert (Yi, n.d.-a).

Abbildung 4: Boxplot über die Verteilung des totalen Niederschlags pro Tag in Mythenquai

Der Boxplot bietet den Vorteil, dass Werte wie der Median und das erste und dritte Quartil direkt abgelesen werden können. Auch wird ein Boxplot meistens dann verwendet, wenn mehrere Kategorien verglichen werden. Dies könnten zum Beispiel mehrere Jahre sein. Da dies in der Abbildung 4 nicht der Fall ist, ist das Histogramm in der Abbildung 3 besser geeignet. Auch ist der Boxplot für die Segler nicht einfach zu lesen und deshalb für mein Zielpublikum ungeeignet.

## Punktewolken/Heatmap

Chart, histogram

Description automatically generated

Abbildung 5: Heatmap: Wasser- vs Lufttemperatur in 2007-2020

Abbildung 6: Punktewolke: Wasser- vs Lufttemperatur in 2007-2020

Abbildung 7: Punktewolke: Wasser- vs Lufttemperatur in 2018

Punktewolken kommen zur Anwendung, wenn man den Zusammenhang zwischen zwei kontinuierliche Variablen darstellen will. Dazu wird auf der x-Achse die eine Variable dargestellt und auf der y-Achse die andere. Für jeden Datenpunkt wird ein Punkt in diesem kartesischen Koordinatensystem erstellt. Durch die Grösse des Punkts kann man noch eine dritte kontinuierliche Variable darstellen, dass setzt aber voraus, dass es nicht viele Punkte gibt (Yi, n.d.-b).

Auch ohne dritte Variable, dürfen nicht zu viele Datenpunkte dargestellt werden. Wenn, wie in Abbildung 4, zu viele Punkte dargestellt werden, verwendet man besser eine Heatmap (Abbildung 5). Bei der Heatmap wird das Koordinatensystem in kleine Rechtecke aufgeteilt, für jedes Rechteck werden die Punkte in diesem Rechteck gezählt und das Rechteck wird entsprechend einer Farbskala eingefärbt (z.B. mehr Punkte => heller) (Yi, n.d.-b).

In der Abbildung 6 ist der Zusammenhang zwischen Wasser- und Lufttemperatur zusehen. Auf der x-Achse ist die Lufttemperatur zusehen, auf der y-Achse die Wassertemperatur. Der Betrachter sieht sofort einen linearen Zusammenhang zwischen der Lufttemperatur und der Wassertemperatur.

# LE2

## Bertin’s sieben visuelle Variablen

Der Kartograph Jacques Bertin hat als erster das Konzept der visuellen Variablen beschrieben. Bei visuellen Variablen handelt es sich um eine visuelle Eigenschaft eines Objekts welche man im Erstellungsprozess verändern kann. In der Datenvisualisierung kann man dies verwenden, um weitere Dimensionen der Daten in der gleichen Visualisierung darzustellen. Dabei korreliert die visuelle Eigenschaft in irgendeiner Weise mit der dargestellten Dimension.

### Grösse

Visuelle Objekte können sich durch ihre Grösse unterscheiden. Die Grösse wird meist durch eine kontinuierliche Variable definiert, sodass diese einen linearen Zusammenhang haben. Dabei ist es für Menschen einfacher diesen Unterschied zu vergleichen und zu erkennen, wenn sich die Grösse nur in einer Dimension verändert (ein Balken wird länger/kürzer). Falls sich die Grösse in zwei (z.B. Kreis) oder drei Richtungen (z.B. Kugel) unterscheidet, haben viele Schwierigkeiten diesen Grössenunterschied zu quantifizieren (The Pennsylvania State University, 2017).

Im Balkendiagramm in der Abbildung 1 stellen die Längen der Balken jeweils den gesamten Niederschlag in diesem Monat dar.

### Form

Die visuelle Variable «Form» wird meistens durch eine kategorielle Variable definiert. Dabei wird die Form des Datenpunkts auf dem kartesischen Koordinatensystem verändert.

Abbildung 8: Bildaufnahme des Flucht- und Rettungsplan des Gebäude 5 Stock 3 der FHNW in Brugg-Windisch

Oftmals werden auf einer Karte verschiedene Symbole verwendet, um verschiedene Arten von Orten zu unterscheiden. In Abbildung 8 sind im Flucht- und Rettungsplan verschiedene Formen verwendet worden um Notausgänge, Handfeuermelder und Defibrillatoren voneinander zu trennen. Bei diesen Formen ist auch gut zu sehen, dass die Form einen Zusammenhang mit der Bedeutung des Punktes hat und nicht z.B. Dreiecke, Kreise und Vierecke verwendet wurden.

Formen wie Balken oder Quadrate überlegen anderen Formen, was das Abschätzen und Vergleichen der visuellen Variable ‘Grösse’ angeht (The Pennsylvania State University, 2017).

### Farbe

Durch die visuelle Variable «Farbe» lassen sich kontinuierliche sowie kategorielle Variablen darstellen. Bei kategoriellen Variablen kann man zum Beispiel auf Karten Flächen (z.B. Länder) markieren. Die Flächen lassen sich aufgrund des Kontrasts der Farben zwischen zwei Flächen auseinanderhalten.

Das Problem dieser Variable ist, dass diese nicht immer von Farbenblinden nicht unterschieden werden kann oder eine Farbskala dieselbe Graustufen hat und somit schwarzweissgedruckt nicht unterschieden werden kann.

In der Abbildung 5 lässt sich durch eine Farbskala die Dichte der Punkte abbilden. Alternativ könnte man hier die Dichte auch durch die Grösse der Rechtecke abbilden.

Durch bestimmte Farbwahlen lassen sich manchmal die psychologischen Assoziationen der Farben mit den Daten verbinden. Zum Beispiel rot für Nein und grün für Ja auf einer Abstimmungskarte.

### Orientierung

Abbildung 9: Windrichtung und Geschwindigkeit in Mythenquai im 2018

Durch die visuelle Variable ‘Orientierung’ lassen sich die Richtungen von Datenpunkten abbilden. Meistens wird auf der Datenvisualisierung das Objekt in die gleiche Richtung ausgerichtet, wie es in der echten Welt auch der Fall ist.

In der Abbildung 9 ist die Windrichtung zu sehen. Die Länge der Balken gibt an, wie oft der Wind in diese Richtung geweht hat. Es wurde ein Polarkoordinatensystem verwendet, auf diesem entspricht die Richtung der Balken auch gleich der Himmelsrichtungen. Dies erleichtert das Interpretieren der Daten. Die Farbskala wurde so gewählt, dass sie auch von farbenblinden Personen interpretiert werden kann.

### Muster

Die visuelle Variable ‘Muster’ definiert sich durch wiederholende Symbole. Auf Landeskarten wird zum Beispiel eine Hecke durch grüne Punkte und Kreise dargestellt (Bundesamt für Landestopografie swisstopo, 2022).

Statt Farben können auch Muster verwendet werden, um verschiedene Flächen auseinander zu halten. Dies ist besonders dann hilfreich, wenn die Visualisierung auch von Farbenblinden gelesen werden muss oder schwarzweiss gedruckt wird.

Mit Muster lassen sich auch kontinuierliche Daten darstellen. Das wird z.B. damit erreicht, dass die kategorielle Variable durch die Dichte von Punkten dargestellt wird (“Visual Variable,” 2022).

## Gestaltgesetze

Die Gestaltgesetze beschreibt mehrere Prinzipien wie der Mensch Objekte wahrnimmt und ordnet. Diese mögen selbstverständlich und von Natur aus gegeben wirken. Dennoch lassen sich dadurch verschiedene Eigenschaften des menschlichen Gehirns verstehen und nutzen.

### Gesetz der Nähe

Das Gesetz der Nähe sagt, dass Objekte mit kleinem Abstand als Gruppen wahrgenommen werden. Dabei kann dieses Gesetz auch auf Musik (z.B. Klopfrhythmen) angewandt werden (Wertheimer, 1923, p. 308).

In der Punktewolke in der Abbildung 7 gruppiert das Auge oben rechts und unten links die Punkte zusammen, da diese nahe beieinander sind.

### Gesetz der Gleichheit

Ähnliche Objekte werden von unserem Gehirn miteinander gruppiert und werden nicht mit unähnlichen vermischt. So gruppieren wir abwechselnde Zeilen von Kreisen und Punkten untereinander als horizontale Linien und nicht als vertikale Linien oder Matrix.

In der Punktewolke in der Abbildung 7 merkt das Auge sofort, dass alle Punkte zum gleichen Phänomen gehören.

### Gesetz der guten Gestalt

Das Gesetz der guten Gestalt oder Gesetz der Prägnanz beschreibt, dass wir simplere Figuren einfacher und schneller wahrnehmen (Weller, 2011). Dies hat zum Beispiel zu Logos wie dieses von Nike geführt.

### Gesetz der guten Fortsetzung

Das Gesetzt der guten Fortsetzung beschreibt folgendes Phänomen: Wenn sich zwei Linien kreuzen, sehen wir nicht zwei aneinander abprallende Linien, welche einen Knick haben. Die Linien folgen somit immer dem einfachsten Weg (Fabian Happacher, 2019).

### Gesetz der Geschlossenheit

Das Gesetz der Geschlossenheit beschreibt die Fähigkeit des Menschen aus nicht geschlossenen Formen geschlossene zu machen. So erkennt unser Hirn bei einem Ring aus Punkten direkt einen geschlossenen Kreis. Auch reicht es bereits nur die Ecken eines Dreiecks zu markieren. Dieses Gesetz wird in der Baubranche verwendet, um Bauvorhaben zu markieren. Dabei wird ein Bauprofil aufgestellt, bei welchem nur die Ecken und Verbindungen zu anderen Ecken markiert werden.

### Gesetz des gemeinsamen Schicksals

Das Gesetz des gemeinsamen Schicksals beschreibt, dass Objekte, welche sich in die gleiche Richtung bewegen als ein Objekt wahrgenommen wird. Zum Beispiel wird eine V-Anordnung von Zugvögel als ein Objekt wahrgenommen (American Psychological Association, n.d.).

# LE3

The main goal with this learning outcome is to make a connection between the data (pre)processing and design decisions.

Data often is not available in the form that is required to use it for data visualizations. Many data sources are in unstructured form and need to be transformed into a format that can be used. During this process, a number of decisions are made how to organize the data. How is the categorical data classified? How were uncertainties in the data handled? How were the classes/categories/binning decided? With this learning goal, you will make connections between the design principles and in the data decisions. Do your visualization design decisions reflect your data decisions? Are your design decisions appropriate for the data type you work with? Why?

Also in connection with other competencies; you will learn different techniques of transforming unstructured into structured data that can be used for visualizations. In parallel, principles of design (in terms of visual hierarchies, layout, typography, color) should be linked to your data decisions. Based on the materials in the 'mediathek' (our media library), you will learn how to design good visualizations in Python. Note that python skills are important in data science, but it is not a requirement for this course. We encourage you to also explore other tools (e.g., R, vega-lite, d3.js). In Python, you will learn how to use different libraries (Matplotlib, Pandas, others) to create visualizations from data.

Things to describe:

1. Wie Daten bereinigen/aufbereiten? und weshalb so? Hat das Auswirken, wenn ich später eine Visualisierung erstellen werde?

* Describe various data sources
* Describe basic obvious things (header, correct type, whats with nan’s, denormalized data, )
* Data Quality Reference Site (What are the 5 metrics of quality data?)

1.1. Wie werden Datentypen identifiziert (categorial, continous, nominal, ordinal)? Wie unterscheiden sich diese

2. Wie werden verschiedene Datenarten (continues, categorial) in einem Plot dargestellt? Welcher Plot ist geeignet

2.1. Wie sollte bin size gewählt werden?

2.2. Gibt es noch andere (komplexere) Visualisierungsarten?

In parallel, principles of design (in terms of visual hierarchies, layout, typography, color) should be linked to your data decisions: TODO connect gestalt principles? Like irgendwie erwähnen?

Sources:

<https://digitalsynopsis.com/wp-content/uploads/2017/12/visual-hierarchy-graphic-design-principles-infographic.jpg>

<https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4>

<https://www.eea.europa.eu/data-and-maps/daviz/learn-more/prepare-data>

# Quellenverzeichnis

American Psychological Association. (n.d.). *Common fate*. Retrieved March 24, 2022, from https://dictionary.apa.org/common-fate

Barbisch, F. (2022). *Gdv*. GitHub. https://github.com/florinbarbisch/gdv

Brönnimann, L. (2021). *Wettermonitor für Wassersportler*. Portrait - (Cde1) Wettermonitor Für Wassersportler - Spaces. https://spaces.technik.fhnw.ch/spaces/wettermonitor-fuer-wassersportler

Bundesamt für Landestopografie swisstopo. (2022). *Zeichenerklärung 2022*. Bundesamt für Landestopografie swisstopo. https://www.swisstopo.admin.ch/de/swisstopo/publikationen.html

Chynał, P., & Sobecki, J. (2016). Eyetracking Evaluation of Different Chart Types Used for Web-Based System Data Visualization. *2016 Third European Network Intelligence Conference (ENIC)*, 159–164. https://doi.org/10.1109/ENIC.2016.031

Fabian Happacher. (2019, January 22). Gesetz der guten Gestalt & der guten Fortsetzung. *VERDINO*. https://verdino.com/blog/gesetz-der-guten-gestalt-und-der-guten-fortsetzung/

Steedle, M. (2010, February 22). *Bar charts vs. Line charts* [Blog]. Axis Insight Blog. https://www.axisgroup.com/data-industry-insights-blog/bar-charts-vs-line-charts

Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, *21*(153), 65–66. JSTOR.

The Pennsylvania State University. (2017, July 13). *Graduated and Proportional Symbol Maps | GEOG 486: Cartography and Visualization*. https://web.archive.org/web/20170713023016/https://www.e-education.psu.edu/geog486/node/1869

Visual variable. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Visual\_variable&oldid=1073462054

Wasserschutzpolizei, Sicherheitsdepartement, & Stadtpolizei. (2021, December 21). *Open Data Zürich—Stadt Zürich*. Messwerte Der Wetterstationen Der Wasserschutzpolizei Zürich. https://data.stadt-zuerich.ch/dataset/sid\_wapo\_wetterstationen

Weller, R. (2011, November 22). *Gestaltgesetze der Wahrnehmung und ihre Bedeutung für das Content Design*. https://www.toushenne.de/design/gestaltgesetze-der-wahrnehmung.html

Wertheimer, M. (1923). *Untersuchungen zur Lehre von der Gestalt (II)*. http://gestalttheory.net/download/Wertheimer1923\_Lehre\_von\_der\_Gestalt.pdf

Yi, M. (n.d.-a). *A Complete Guide to Box Plots* [Guide]. Chartio. Retrieved March 17, 2022, from https://chartio.com/learn/charts/box-plot-complete-guide/

Yi, M. (n.d.-b). *A Complete Guide to Scatter Plots* [Guide]. Chartio. Retrieved March 17, 2022, from https://chartio.com/learn/charts/what-is-a-scatter-plot/

# Anhangverzeichnis

# Anhang

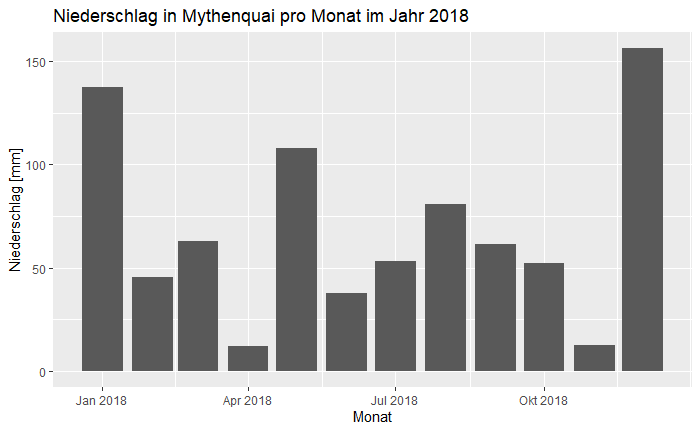
Abbildung 1: Balkendiagramm über den Niederschlag in Mythenquai pro Monat im Jahr 2018

Abbildung 2: Liniendiagramm über den Verlauf der Lufttemperatur in Mythenquai am ersten Januar 2018

Chart, line chart

Description automatically generated

Abbildung 3: Histogramm über die Verteilung des totalen Niederschlags pro Tag in Mythenquai



Abbildung 4: Punktewolke: Wasser- vs Lufttemperatur in 2007-2020

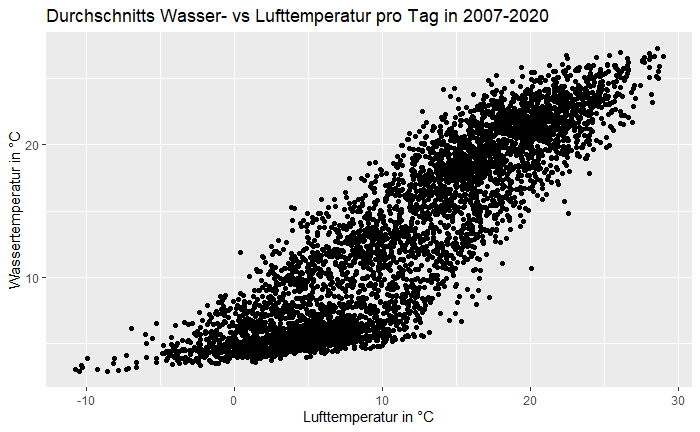


Abbildung 5: Heatmap: Wasser- vs Lufttemperatur in 2007-202

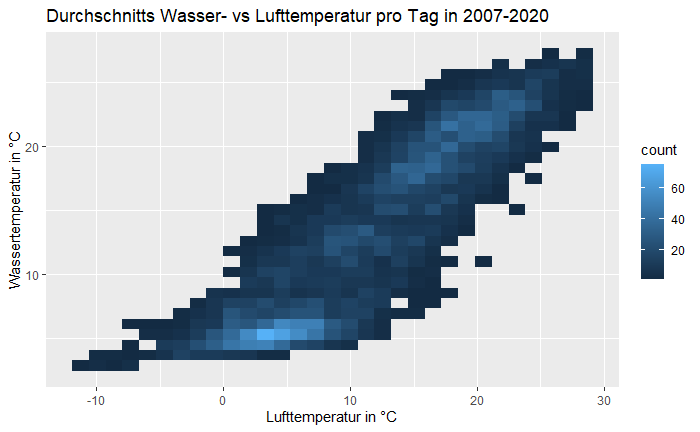


Abbildung 6: Punktewolke: Wasser- vs Lufttemperatur in 2018

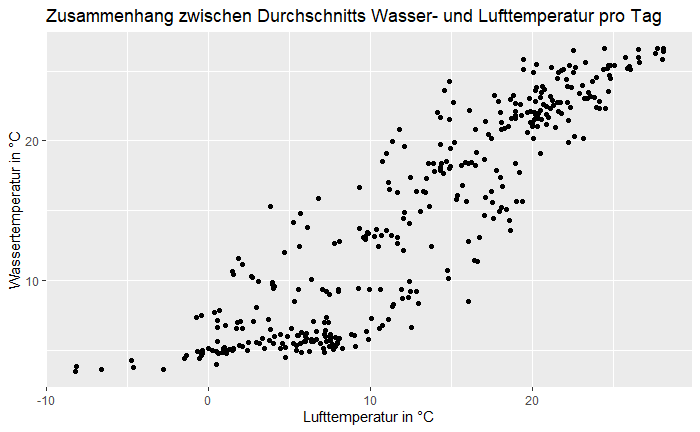


Abbildung 8: Flucht- und Rettungsplan des Gebäude 5 Stock 3 der FHNW in Brugg-WindischDiagram, engineering drawing

Description automatically generated