Grundlagen Datenvisualisierungen

Bericht

A small boat in a body of water

Description automatically generated

Student: Florin Barbisch

Studiengang: FHNW, Bsc Data Science

Fachexperten: Arzu Cöltekin

Martin Lacayo

Datum: 12.10.2022

Table of Contents

[2 LE1 3](#_Toc116480112)

[2.1 Balkendiagramm 3](#_Toc116480113)

[2.2 Liniendiagramm 3](#_Toc116480114)

[2.3 Histogramm 4](#_Toc116480115)

[2.4 Boxplot 4](#_Toc116480116)

[2.5 Punktewolken/Heatmap 4](#_Toc116480117)

[3 LE2 5](#_Toc116480118)

[3.1 Bertin’s sieben visuelle Variablen 5](#_Toc116480119)

[3.2 Gestaltgesetze 6](#_Toc116480120)

[4 LE3 7](#_Toc116480121)

[4.1 Data cleansing 7](#_Toc116480122)

[4.2 Qualitätskriterien 7](#_Toc116480123)

[4.3 Datentypen 8](#_Toc116480124)

[5 LE4 9](#_Toc116480125)

[5.1 Grammar of Graphics 9](#_Toc116480126)

[6 LE5 11](#_Toc116480127)

[6.1 Usability-Tests 11](#_Toc116480128)

[7 Quellenverzeichnis 13](#_Toc116480129)

[8 Anhang 16](#_Toc116480130)

[8.1 Abbildung 1: Balkendiagramm über den Niederschlag in Mythenquai pro Monat im Jahr 2018 16](#_Toc116480131)

[8.2 Abbildung 2: Liniendiagramm über den Niederschlag in Mythenquai pro Monat im Jahr 2018 16](#_Toc116480132)

[8.3 Abbildung 3: Histogramm über die Verteilung des totalen Niederschlags pro Tag in Mythenquai 17](#_Toc116480133)

[8.4 Abbildung 4: Boxplot über die Verteilung des totalen Niederschlags pro Tag in Mythenquai 17](#_Toc116480134)

[8.5 Abbildung 5: Heatmap: Wasser- vs Luft-temperatur in 2007-2020 18](#_Toc116480135)

[8.6 Abbildung 6: Punktewolke: Wasser- vs Luft-temperatur in 2007-2020 18](#_Toc116480136)

[8.7 Abbildung 7: Punktewolke: Wasser- vs Lufttemperatur in 2018 19](#_Toc116480137)

[8.8 Abbildung 8: Bildaufnahme des Flucht- und Rettungsplan des Gebäude 5 Stock 3 der FHNW in Brugg-Windisch 20](#_Toc116480138)

[8.9 Abbildung 9: Windrichtung und Geschwindigkeit in Mythenquai im 2018 21](#_Toc116480139)

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

In der **Error! Reference source not found.** 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 erkennen, wo sich der Datenpunkt befindet. Deshalb ist für diese Visualisierung das Balkendiagramm besser geeignet.

## Histogramm

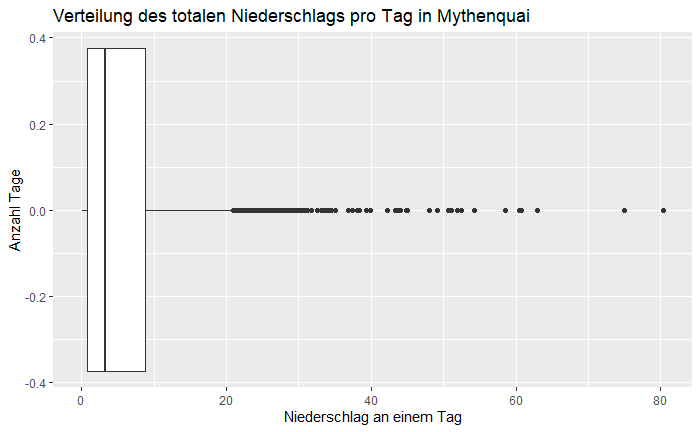
In einem Histogramm kann eine Verteilung einer kontinuierlichen Variable dargestellt werden. Es werden ähnlich wie im Balkendiagramm Balken gezeichnet, die Höhe der Balken entspricht aber der Anzahl Werte, welche im Intervall des Balken vorkommen. Die x-Achse muss sortiert sein, sonst lässt sich keine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

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.

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

## Chart, histogram Description automatically generatedPunktewolken/Heatmap

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).

Abbildung : Heatmap: Wasser- vs Luft-temperatur in 2007-2020

Abbildung : Punktewolke: Wasser- vs Luft-temperatur in 2007-2020

Abbildung : Punktewolke: Wasser- vs Lufttemperatur in 2018

Auch ohne dritte Variable, dürfen nicht zu viele Datenpunkte dargestellt werden. Wenn, wie in Abbildung 6, 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 7 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 : 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 : 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

## Data cleansing

Daten kommen meistens von verschiedenen Datenquellen und damit in unterschiedlicher Datenqualität. Um gute Datenvisualisierungen zu machen, muss die Datenqualität sichergestellt werden. Dazu müssen die Daten bereinigt werden. Dies wird im Data Management ‘Data cleansing’ genannt (Tableau, n.d.).

Verschiedene Datenquellen kommen in unterschiedlicher Qualität. So sind die CSV-Dateien der Wetterstation von einem Computer generiert worden und haben deshalb eine klare Struktur. In der Temperaturspalte sind zum Beispiel nur Zahlen und keine Buchstaben. Auch ist bei der Zeitspalte die Zeitzone schon definiert. Wenn die Daten aus einer Datenbank gelesen werden, ist es sinnvoll die Daten ein Sternschema zu bringen und dabei zu denormalisieren (Microsoft Corporation, 2022). Im Gegensatz zu meinen CSV-Dateien sind bei manuell erfassten Daten mehr Arbeit notwendig, da der Mensch fehleranfällig ist.

## Qualitätskriterien

Um bei Data Cleansing die Qualität sicherzustellen, werden folgende 5 Qualitätskriterien verwendet.

### Validität

Valide Daten beschreiben das Konzept, dass die Daten sehr wahrscheinlich der Realität entsprechen. Dies wird meistens durch Regeln sichergestellt (Rachel Levy Sarfin, 2021). Zum Beispiel darf die Luftfeuchtigkeit der Wetterdaten von Mythenquai nie unter 0% liegen (aber sehr wohl über 100% (“Luftfeuchtigkeit,” 2022)). Oder die Windrichtung muss zwischen 0° und 360° liegen. Es gibt noch weitere Bedingungen wie, dass der Zeitstempel nie leer sein darf und einmalig sein muss. In den Daten war zu Beginn bei der Zeitumstellung vom Wechsel von Sommer- zu Winterzeit die Zeitstempel zwischen 02:00 und 03:00 nicht eindeutig (keine Zeitzone). Wäre dies nicht so würden in einem Liniendiagramm des Tagesverlauf zwischen 02:00 und 03:00 zwei Linien erscheinen oder ein gerader Strich, weil es keinen Daten gibt.

### Genauigkeit

Die Genauigkeit von Daten besteht aus zwei Kriterien (*ISO 5725-1*, 1994):

* Die Präzision beschreibt, dass einzelne Datenpunkte korrekt sind. Dies ist vor allem bei kontinuierlichen Daten nicht einfach zu überprüfen. Bei kategoriellen Daten wie Adressen können externe Datenquellen zur Überprüfung dazu gezogen werden.
* Der Bias beschreibt eine systematische Abweichung aller Werte von den korrekten Werten.

### Vollständigkeit

Daten sind vollständig, wenn alle benötigten Daten vorhanden sind (Rachel Levy Sarfin, 2021). Dies ist relativ einfach zu überprüfen aber meistens unmöglich dies im Nachhinein zu beheben.

In den Wetterdaten, fehlen an diversen Tagen Messwerte. So wurde am 08.04.2019 kein einziger Messwert erfasst. Hätte es an diesem Tag viel geregnet, würde der Balken vom April in der Abbildung 1 viel kleiner ausfallen. Aufgrund Bauarbeiten fehlen seit Anfangs 2020 die Wassertemperatur, Seespiegel, Niederschlag und Strahlung (Wasserschutzpolizei et al., 2021).

### Konsistenz

Konsistenz sagt aus, dass die Daten über mehrere Systeme gleich sind (Rachel Levy Sarfin, 2021). Dies lässe sich bei der Lufttemperatur sicherstellen, indem die sie mit der Lufttemperatur von anderen Wetterstationen verglicht.

### Einheitlichkeit

Die Einheitlichkeit zeichnet sich dadurch aus, dass alle Daten die gleiche Einheit haben (Elgabry, 2019). In den Wetterdaten wurde keine Zeitzone angegeben. Wenn Daten aus dem Sommer als Winterzeit interpretiert werden, würde der Tagesverlauf der Temperatur auf einem Liniendiagramm um eine Stunde verschoben dargestellt. Diese Uneinheitlichkeit wurde auf Anfrage von OpenDataZürich behoben.

## Datentypen

Um bei Daten zu entscheiden welche Visualisierung passt, hilft es den Datentyp zu bestimmen. Es gibt im Wesentlichen vier davon.

### Quantitative Daten

Als quantitative Daten wird alles bezeichnet, was eine Zahl ist (Wilke, 2019, Chapter 2.1). Diese besitzen immer eine natürliche Reihenfolge. Zu den quantitativen Daten gehören zum Beispiel Körpergrössen oder Anzahl Personen. Man unterscheidet weiter zwischen diskreten und kontinuierlichen Daten.

#### kontinuierlichen Daten

Die kontinuierlichen Daten zeichnen sich dadurch aus, dass es zwischen einem beleibeigen Intervall eine unendliche Anzahl an Zwischenwerte gibt (Wilke, 2019, Chapter 2.1). Dies ist bei vielen Messgrössen der Fall.

In den Wetterdaten sind alle Messwerte kontinuierlich. Streng genommen sind die Daten als diskrete Daten abgespeichert. So wird z.B. die Lufttemperatur auf einen Zehntel gerundet und es gibt damit im Intervall 10 bis 20 nun 100 abzählbare Werte und nicht mehr unendlich viele. Trotz dieser Rundung wird diese Messgrösse immer noch als kontinuierliche behandelt.

Um die Verteilung einer kontinuierlichen Variable darzustellen, wird ein Histogramm verwendet werden, wie es im Kapitel ‘2.3 Histogramm’ und in der Abbildung 3 beschrieben ist. Ein Barplot ist nicht geeignet, da bei diesem die Balken nach der Häufigkeit sortiert werden und dann der Zusammenhang zwischen den Werten weg ist.

#### Diskrete Daten

Diskrete Grössen sind alle Daten, welche in einem beliebigen Intervall eine abzählbare Anzahl an Werten haben. So ist zum Beispiel die Anzahl Personen zwischen 10 und 20 abzählbar (Wilke, 2019, Chapter 2.1). Es gibt 9 verschiedene Werte dazwischen.

Für die Abbildung 1 wurden die Zeitstempel zu Monaten zusammengefasst und so eine diskrete Grösse gebildet. Würde man dies nicht machen, hätte man mehr Datenpunkte und Balken in der Abbildung.

### Qualitative Daten

Als qualitative Daten werden Daten bezeichnet, welche keine Zahlen sind. Diese können eine natürliche Reihenfolge haben oder nicht (Wilke, 2019, Chapter 2.1). Zu qualitativen Daten gehören zum Beispiel Kleidergrössen (S, M, L), Wochentage oder Postleitzahlen. Weiter wird zwischen ordinalen und nominalen Grössen unterschieden.

#### Ordinale Daten

Ordinale Daten haben immer eine natürliche Reihenfolge (Wilke, 2019, Chapter 2.1). Dazu gehören zum Beispiel Kleidergrössen.

Bei den Wetterdaten gibt es per se keine ordinalen Daten. Für die Abbildung 9 wurde die Windrichtung in die Himmelsrichtungen Norden, Osten, Süden und Westen unterteilen und so ordinale Grösse gebildet. Damit wird die Windrichtung vereinfacht und ist einfacher zu lesen. Wenn man die Windrichtung in Grad lesen muss, ist das schwieriger, weil man im Alltag oftmals in Himmelsrichtungen und nicht in Grad denkt.

#### Nominale Daten

Nominale Daten haben keine natürliche Reihenfolge (Wilke, 2019, p. 2.1). Dazu gehören Automarken oder Postleitzahlen. Postleitzahlen können zwar als Zahlen gelesen werden, dennoch sind es keine Zahlen, weil zum Beispiel die Addition überhaupt keinen Sinn macht (meh, 2016).

In den Wetterdaten ist nur der Name der Wetterstation (Mythenquai oder Tiefenbrunnen) eine nominale Grösse.

# LE4

## Image for postGrammar of Graphics

Damit man Graphiken nicht durch Trail and Error erstellt, strukturiert man den Prozess des Erstellens der Graphiken. Ein solches Konzept, welches wir hier ansehen heisst «Grammar of Graphics». Es wurde von Leland Wilkinson im gleichnamigen Buch vorgestellt (Leland Wilkinson, 2005), basiert auf Bertin’s sieben Variablen, welche im Kapitel 3.1 beschrieben sind, und durch die Library ggplo2 popularisiert wurde (Wickham et al., 2022). «Grammar of Graphics» teilt Graphiken in verschiedene Ebenen auf. Eine Übersicht dieser ist in der Abbildung 10 ersichtlich (Sarkar, 2018). In den nachfolgenden Unterkapitel werden diese sieben verschiedenen Ebenen vorgestellt und auf die Windrose in der Abbildung 9 angewendet.

Abbildung : Hierarchie der Grammar of Graphics

### Daten

Die Daten bilden das Fundament jeder Grafik und stellen die unterste Ebene der Grammar of Graphics Pyramide dar (Sarkar, 2018). In meinem Fall verwende ich die Windgeschwindigkeit und Windrichtung von den Daten des Wettermonitors. Die Daten des Wettermonitors sind empirischer Natur und kommen aus einem Dataframe.

### Aesthetics

Aesthetics oder das Mapping der Aesthetics beschreiben das Zuweisen der Variablen der Daten zu einer Achse (meisten der x- und y-Achse) oder der Farbe, Grösse, Form etc. (Soltoff, 2019). Im Fall der Windrose wird der x-Achse die Windrichtung, der y-Achse die Anzahl der Messungen und der Farbeskala die Windgeschwindigkeit zugewiesen. Wenn man die Zuweisung der x- und y-Achse vertauschen würde, hätte man später im polaren Koordinatensysstem Ringe, was sehr verwirrend wäre.

### Scale

Abbildung : Windrose im kartesischen Koordinatensystem

Mithilfe von Skalen werden die Daten auf eine Grösse abgebildet die für die Aesthetics sinnvoll sind. Diese Grösse wird auch verwendet, um die Grafik zu rendern (Soltoff, 2019). So wird bei der Windrose die x-Achse in acht Himmelsrichtungen aufgeteilt und die Windgeschwindigkeit in einen Farbwert konvertiert. Für den Farbwert ist es besser eine Farbskala zu nehmen, die von hell zu dunkel verläuft statt die Regenbogenfarben, da Personen Farben besser in eine Reihenfolge bringen können, wenn diese durch die Helligkeit gegeben wird (Golebiowska & Coltekin, 2022). Eine häufige Abbildung bei Achsen ist das Logarithmieren. Das macht bei meinen Daten aber weniger Sinn, da die Daten nicht eine grosse Spannweite haben.

### Geometric objects

Mit geometrischen Objekten wird die Art der Graphik beschrieben. Ein geometrisches Objekt kann immer nur bestimmte Aesthetics abbilden. Es wird zwischen drei Arten von geometrischen Objekten unterschieden, diese werden durch ihre Anzahl Dimensionen identifiziert. Null Dimensionen repräsentiert einen Punkt oder Text. Eine Dimension ist eine Linie und mit zwei Dimensionen werden Polygone oder Intervalle (Wickham, 2010). Die Säulen in der Windrose sind hier das geometrische Objekt mit *einer* Dimension (die Breite der Säulen bildet hier keine Variable ab). Die Säulen, wie sie in der Abbildung 11 gut zu sehen sind, haben hier die folgenden Aesthetics: die Höhe (Anzahl Messungen) und die Farbe (Windgeschwindigkeit).

### Statistics

Bei statistischen Transformationen werden die Daten transformiert. Oft geschieht dies durch Zusammenfassen der Daten (Reduktion von mehreren Datenpunkten auf einzelne Fakten). Solche Transformationen beinhalten zum Beispiel: Anzahl, Durchschnitt, Median, Standardabweichung etc (Soltoff, 2019). Bei Säulendiagrammen stellt oft die Höhe die Anzahl dar. Bei Boxplot sind gerade mehrere statistische Transformationen notwendig. Es wird neben dem Median das erste und dritte Quartil sowie das Minimum und Maximum berechnet. In einem Punktediagramm muss keine statistische Transformation vorgenommen werden, wenn nur 2 Variablen dargestellt werden). Im Beispiel der Windrose muss nur die Anzahl berechnet werden, gruppiert nach Geschwindigkeit und Windrichtung. Statt der Anzahl hätte man auch die relative Anzahl verwenden können. Dann wäre in der Abbildung zwar die Verteilung der Windgeschwindigkeiten einfacher ablesbar, da auch Balken mit wenigen Punkten im Total gleich lange wären wie die anderen. Aber es wäre nicht mehr möglich, herauszufinden, dass in Richtung Nordosten der Wind am meisten weht.

### Facets

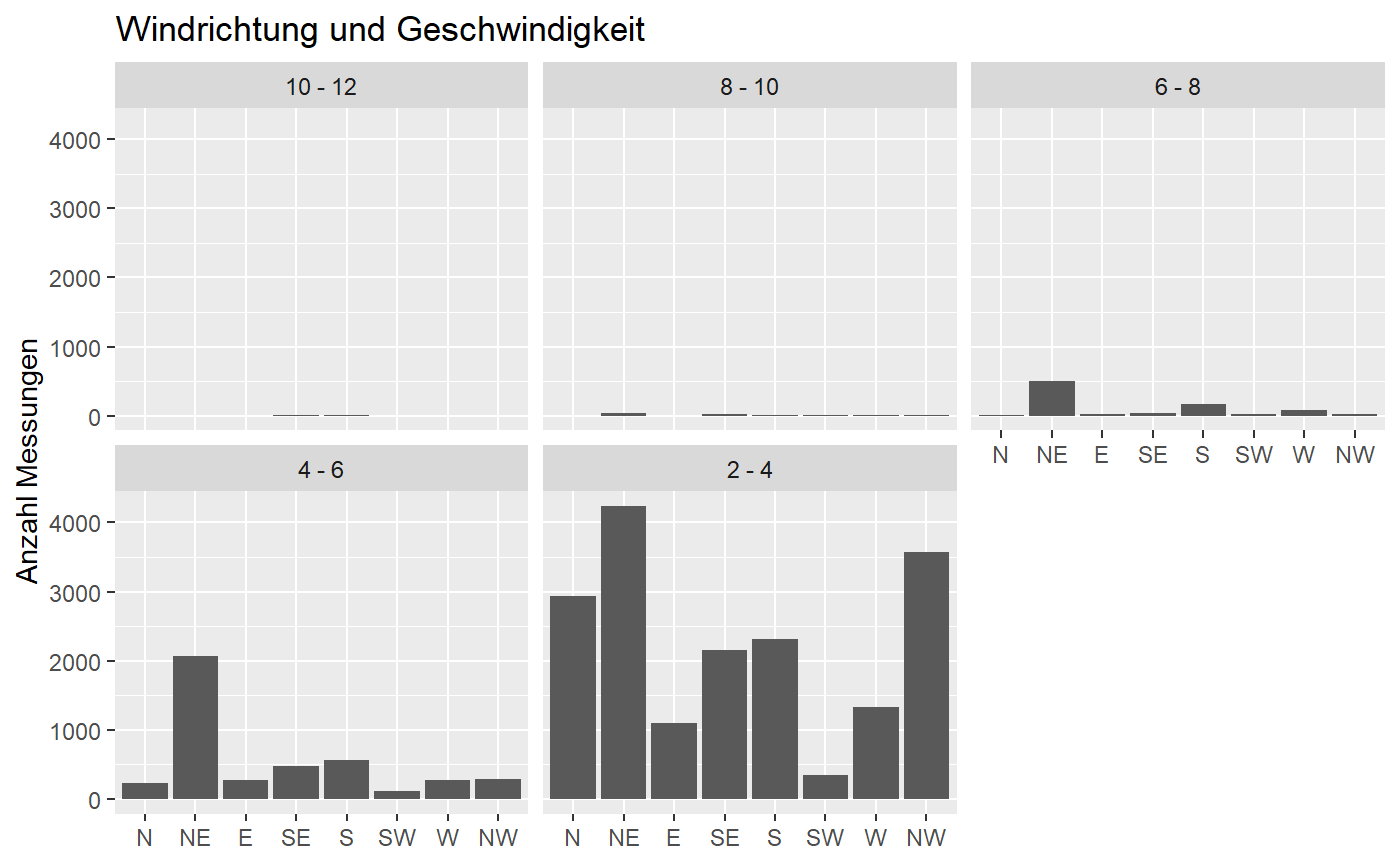
Manchmal möchte man untersuchen, ob Muster bei verschiedenen Kategorien gleich auftreten oder es innerhalb gewisser Kategorien Muster gibt. Dazu ist es hilfreich, wenn man für jeden Wert innerhalb dieser Kategorie eine eigene Grafik hat. Genau das kann mit Hilfe von Facetten erreicht. Dabei werden die Daten nach jeder Kategorie aufgeteilt und bilden sozusagen einen neuen Datensatz für eine Teilgrafik (Soltoff, 2019). Die Facetten können als Spalten, als Zeilen oder, wenn zwei Variablen für die Aufteilung der Grafik verwendet werden, auch als Raster. In der Abbildung 12 ist die Geschwindigkeit als Facette abgebildet. Im Vergleich zur Abbildung 11 werden hier keine neuen Informationen dargestellt. Aber indem man bei der Abbildung 12 für jede Geschwindigkeit eine neue Abbildung erstellt, beginnen die Balken immer bei der Nulllinie im Gegensatz zur Abbildung 11, bei welcher die Balken gestapelt werden. Damit kann die Höhe der Balken von verschiedenen Himmelsrichtung besser verglichen werden.

Abbildung : Windrichtung mit Windgeschwindigkeit als Facetten

### Coordinate System

Chart, radar chart

Description automatically generatedDas Koordinatensystem bestimmt, wie und vor allem wo Objekte auf der Graphik gezeichnet werden. Neben den geometrischen Objekten werden insbesondere auch die Skalen durch das Koordinatensystem abgebildet. Graphiken besitzen meistens nur 2 Koordinatenachsen, könnten aber auch mehr als 2 Koordinatenachsen besitzen. Wie diese Objekte jeweils abgebildet werden, wird durch das Koordinatensystem bestummen. Hier gibt es mehrere Arten davon (Soltoff, 2019). Das Bekannteste ist das kartesische Koordinatensystem so wie es zum Beispiel in der Abbildung 12 verwendet wird. Dies kann noch weiter konfiguriert werden, indem das Seitenverhältnisse der Achsen konfiguriert wird. Bei der Windrose möchte man die Himmelsrichtungen so darstellen, dass sie in einem Kreis angeordnet sind. Das macht es einfacher die Grafik zu lesen, weil die Himmelsrichtungen mit gleich wie bei einem Kompass angeordnet ist. Dies wird durch ein polares Koordinatensystem wie in der Abbildung 9 erreicht.

Abbildung 9: Windrichtung und Geschwindigkeit in Mythenquai im 2018

# LE5

Bei fertigen Visualisierungen stellt sich oft die Frage, ob die Visualisierung gut ist oder nicht, respektive wie gut sie ist. Um die Qualität der Visualisierung zu messen kann man verschiedene Evaluierungsmethoden verwenden Ich habe mich Usability-Tests entschieden, da ich damit Usability-Probleme in meinen Visualisierungen entdecken kann (Lee, 2018) und möchte. Ich könnte anstelle einer qualitativen Forschungsmethode wie Usability-Tests eine quantitative Forschungsmethode wählen. Dadurch ist zwar die Gefahr kleiner, dass die Ergebnisse durch meine eigene Meinung beeinflusst wird. Aber es wird damit nicht möglich sein einen tieferen Einblick in das Vorgehen der Personen beim Beantworten der Fragen zu erlangen. Deshalb habe ich mich für eine Usability-Test entschieden, welche ich im folgenden Kapitel durchführen werde.

## Usability-Tests

Durch Usability-Tests kann die Visualisierung verbessert werden. Anschliessend kann erneut ein Usability-Tests durchgeführt werden und so durch einen Iterativen Prozess die Visualisierung perfektioniert werden. Usability-Tests haben im allgemeinen drei Ziele: 1. sie versuchen Probleme mit der Visualisierung zu entdecken, 2. sie sollen Möglichkeiten aufzeigen, wie man die Visualisierung verbessern kann und 3. man lernt während dem Durchführen neues über das Verhalten und die Vorlieben der Benutzer (Kate Moran, 2019).

Bei Usability-Tests gibt die moderierende Person der teilnehmenden Person Aufgaben, die diese ausführen muss. Die teilnehmende Person gibt dabei mündliches Feedback und die moderierende Person beobachtet die teilnehmende Person. Die moderierende Person stellt meistens Fragen um noch mehr Informationen von der anderen Person zu erhalten (Kate Moran, 2019).

### Aufgaben

Mich interessiert ob man aus meinen bisherigen Visualisierungen Informationen ablesen kann und ob die teilnehmenden Personen noch Probleme aufdecken. Die Aufgaben bestehen darin, die Fragen zur jeweiligen Visualisierung zu beantworten. Den Teilnehmenden wird gesagt, dass sie die Grafik in einem Zeitungsartikel sehen und folgenden Fragen beantworten müssen.

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

* Wann hat es am meisten geregnet?
* Wann war es am trockensten?
* Wieviel hat es im Mai 2018 geregnet?

Abbildung 9: Windrichtung und Geschwindigkeit in Mythenquai im 2018:

* Aus welcher Richtung kommt der Wind am meisten?
* Wievielmal wurde Wind aus der Richtung Nordwesten gemessen?
* Welche Windgeschwindigkeit wurde in Richtung Norden am zweitmeisten gemessen?

### Teilnehmende

Da Usability-Tests für die moderierende Person sowie die Teilnehmenden aufwändig sind, möchte man nicht zu viele Personen testen. Da es sich hier um einen qualitativen Test handelt, reichen 5 Personen (Jakob Nielsen, 2000). Es ist wichtig, dass die Teilnehmenden mögliche Nutzer sind (Kate Moran, 2019). Da die Aufgabe vorsieht, dass die Visualisierungen in einer Zeitung erscheinen, kommen als Teilnehmenden alle Personen infrage, die (Deutsch) Lesen können. Um nicht zu technisch versierte Personen zu befragen, habe ich auf meine Mitstudierenden verzichtet und stattdessen Personen aus meinem familiären Umfeld befragt.

### Durchführung

Wie vorhin erwähnt habe ich den Test an 5 Personen aus meine familiären Umfeld einzeln durchgeführt. Ich habe den Personen jeweils erklärt, dass sie die Grafik in der Zeitung gefunden haben und die Aufgaben dazu beantworten müssen. Des Weiteren habe ich sie darum gebeten, ihren laut zu denken, sodass ich verstehen kann was bei den Teilnehmenden abläuft (Nielsen, 2012). Ich habe die Visualisierung nacheinander gegeben, in der gleichen Reihenfolge wie oben beschrieben. Am Ende jeder Visualisierung habe ich die Teilnehmenden gefragt, ob ihnen noch etwas an den Visualisierungen auffallen würde oder sonst etwas stört.

### Auswertung der Usability-Tests

#### Abbildung 1

Die erste Frage «Wann hat es am meisten geregnet?» konnten alle beantworten, hilfreich waren hier die horizontalen Gitternetzlinien. Eine Personen hat darauf hingewiesen, dass Niederschlag nicht gleich Regen ist und somit die Frage nicht ganz korrekt beantwortbar ist. Zwei haben sich weiter gefragt, wieso es im Winter so viel geregnet hat und ob das eventuell die Schneehöhe ist. Beim Interpretieren der Grafik haben 4 Personen geäussert, dass sie annehmen, dass die Monat sortiert sind. Hier wäre also eine Beschriftung von jedem Balken sinnvoll.

Die zweite Frage «Wann war es am trockensten?» wurden immer zwei Monate (April und November) genannt, da die Balken etwa gleich hoch sind. Tatsächlich hat es im April auch nur 0.5mm weniger geregnet. Es ist daher auch zu erwarten, dass bei dieser Frage nicht nur ein Monat kommt.

Die dritte Frage «Wieviel hat es im Mai 2018 geregnet?» war eine Schätzfrage und wurde sehr gut geschätzt. Einer Person war beim Lesen der Frage zuerst nicht klar ob nach Tagen oder Millimeter gefragt wird. Zwei Personen haben sich beim Lesen der Visualisierung gefragt, wo auf der Welt Mythenquai ist. Ich könnte hier im Untertitel der Grafik schreiben, dass es sich um Wetterdaten der Messtation Mythenquai in Zürich handelt, dann wäre diese Frage auch geklärt.

Die Erkenntnis ist, dass das Balkendiagramm gut ist um den Niederschlag abzulesen und zwischen einzelnen Monaten zu vergleichen, sofern die Unterschiede nicht allzu klein sind.

#### Abbildung 9

Die Grafik war für alle Teilnehmenden nicht ganz einfach zu lesen. Erstens war nicht allen klar, dass mit Windrichtung nicht das gleiche gemeint, wie die Richtung aus der der Wind kommt. Zweitens war es für alle schwer die y-Achsenbeschriftung am linken Rand auf die Gitternetzlinien (die ja Kreise sind) zu übertragen, auch weil die 8000 Markierung fehlt. Eine Person interpretierte die Messwerten in Richtung Süden als Minuswerte. Dies kann man beheben, indem die Achsenbeschriftung in die Mitte (auf die Nordachse) setzt und noch die 8000 Linie markiert. Bei den Balken dachte jemand zuerst, dass diese gestapelt sind. Die Himmelsrichtungen wurden trotz Abkürzungen von allen problemlos erkannt.

Die Frage «Aus welcher Richtung kommt der Wind am meisten?» konnten 3 richtig beantworten. Die anderen beiden Personen haben die gegenüberliegende Himmelsrichtung angegeben. Eine Person hatte Mühe zu erkennen, dass mit *meisten* der grösste Wert von *Anzahl Messungen* gemeint ist.

Bei der zweiten Frage «Wievielmal wurde Wind aus der Richtung Nordwesten gemessen?» war wieder das gleiche Problem wie bei der ersten Frage: 2 haben die Anzahl an der gegenüberliegenden Achse abgelesen. Die restlichen drei Personen konnten die Anzahl auf 100-300 Messungen genau abschätzen. Eine genauere Abschätzung war nicht möglich, da der Balken relativ klein ist und es nicht mehr Gitternetzlinien gibt.

Die letzte Frage «Welche Windgeschwindigkeit wurde in Richtung Norden am zweitmeisten gemessen?» war für alle nicht ganz einfach zu lösen, es konnten sie dann jedoch alle korrekt beantworten. Den Teilnehmenden ist auch aufgefallen, dass in alle Richtungen die Windgeschwindigkeit 4-6 m/s die zweitmeiste ist. Zwei Teilnehmenden ist aufgefallen, dass in der Legende die Windgeschwindigkeit 0-2 m/s fehlt.

Ich habe gemerkt, dass die Windrose diversen Verbesserung bedarf und eher schwieriger zu interpretieren ist aber zum Abbilden von Himmelsrichtungen geeignet.

# Quellenverzeichnis

*Accuracy (trueness and precision) of measurement methods and results—Part 1: General principles and definitions*. (1994). ISO. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/01/18/11833.html

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

Elgabry, O. (2019, March 2). *The Ultimate Guide to Data Cleaning*. Medium. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

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/

Golebiowska, I. M., & Coltekin, A. (2022). Rainbow Dash: Intuitiveness, Interpretability and Memorability of the Rainbow Color Scheme in Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *28*(7), 2722–2733. https://doi.org/10.1109/TVCG.2020.3035823

Jakob Nielsen. (2000, March 18). *Why You Only Need to Test with 5 Users*. Nielsen Norman Group. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Kate Moran. (2019, December 1). *Usability Testing 101*. Nielsen Norman Group. https://www.nngroup.com/articles/usability-testing-101/

Lee, S. (2018, February 15). *User Research vs. Usability Testing*. Medium. https://build.diligent.com/user-research-usability-testing-e223027e4c2d

Leland Wilkinson. (2005). *The Grammar of Graphics*. Springer-Verlag. https://doi.org/10.1007/0-387-28695-0

Luftfeuchtigkeit. (2022). In *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Luftfeuchtigkeit&oldid=221865584

meh. (2016). *Categorical or Quantitative?* https://stats.stackexchange.com/q/203082

Microsoft Corporation. (2022, April 7). *Understand star schema and the importance for Power BI - Power BI*. https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

Nielsen, J. (2012, January 15). *Thinking Aloud: The #1 Usability Tool*. Nielsen Norman Group. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

Rachel Levy Sarfin. (2021, April 30). Data Quality Dimensions: How Do You Measure Up? *Precisely*. https://www.precisely.com/blog/data-quality/data-quality-dimensions-measure

Sarkar, D. (DJ). (2018, September 13). *A Comprehensive Guide to the Grammar of Graphics for Effective Visualization of Multi-dimensional…*. Medium. https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149

Soltoff, B. (2019, March 1). *The grammar of graphics*. Computing for the Social Sciences. https://cfss.uchicago.edu/notes/grammar-of-graphics/

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.

Tableau. (n.d.). Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data. *Tableau*. Retrieved April 7, 2022, from https://www.tableau.com/learn/articles/what-is-data-cleaning

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). *Messwerte der Wetterstationen der Wasserschutzpolizei 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

Wickham, H. (2010). A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics*, *19*(1), 3–28. https://doi.org/10.1198/jcgs.2009.07098

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & RStudio. (2022). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (3.3.6). https://CRAN.R-project.org/package=ggplot2

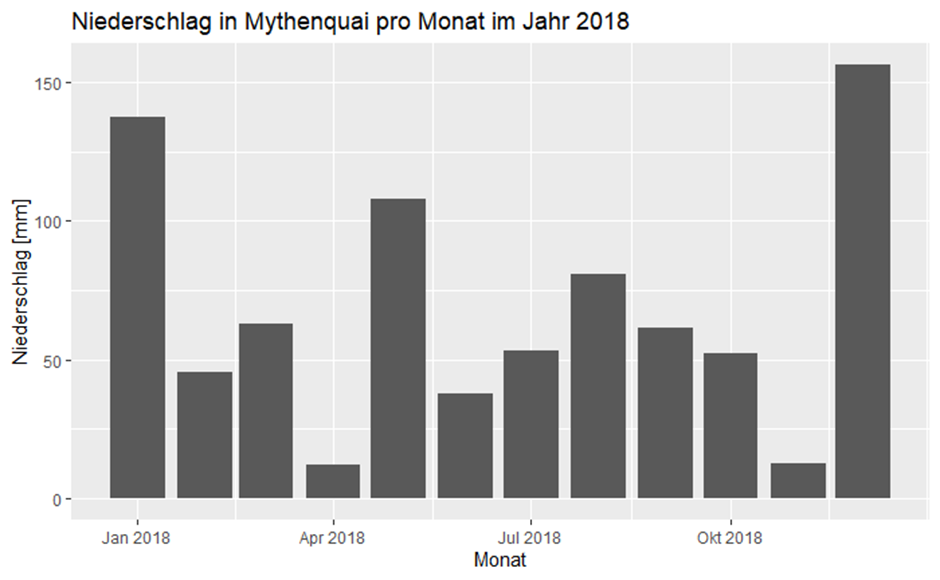
Wilke, C. O. (2019). *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O’Reilly Media, Inc.

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/

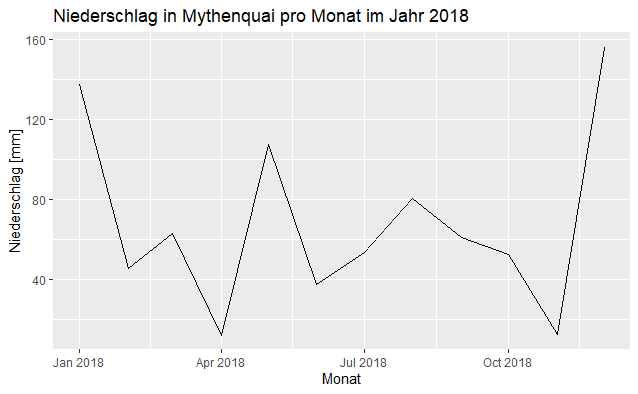
Yi, M. (n.d.-b). *A Complete Guide to Scatter Plots* [Guide]. Chartio. Retrieved March 17, 2022, from https://chartio.com/learn/charts/what-is-a-scatter-plot/

# Anhang

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



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

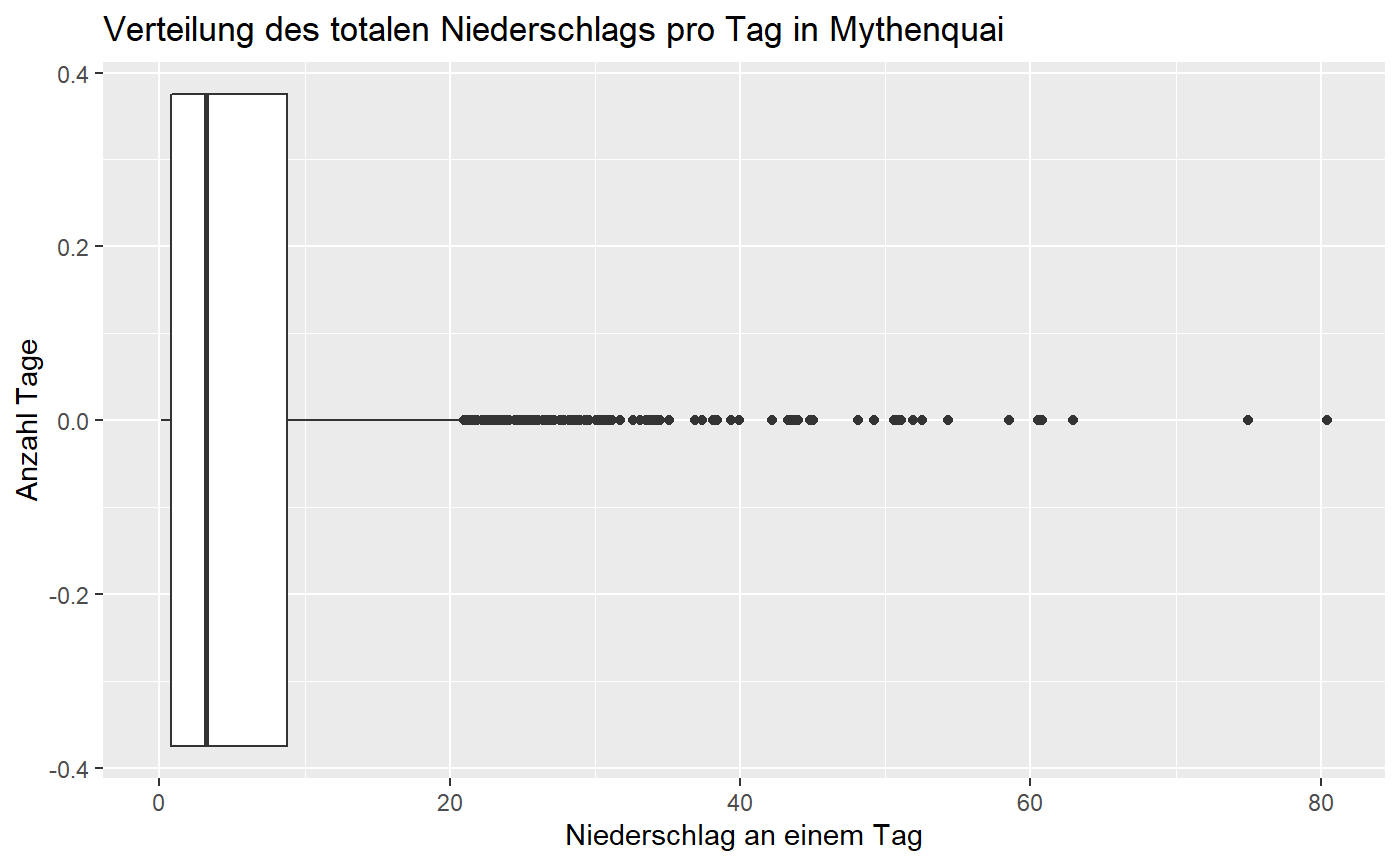


## Abbildung 3: Histogramm über die Verteilung des totalen Niederschlags pro Tag in Mythenquai

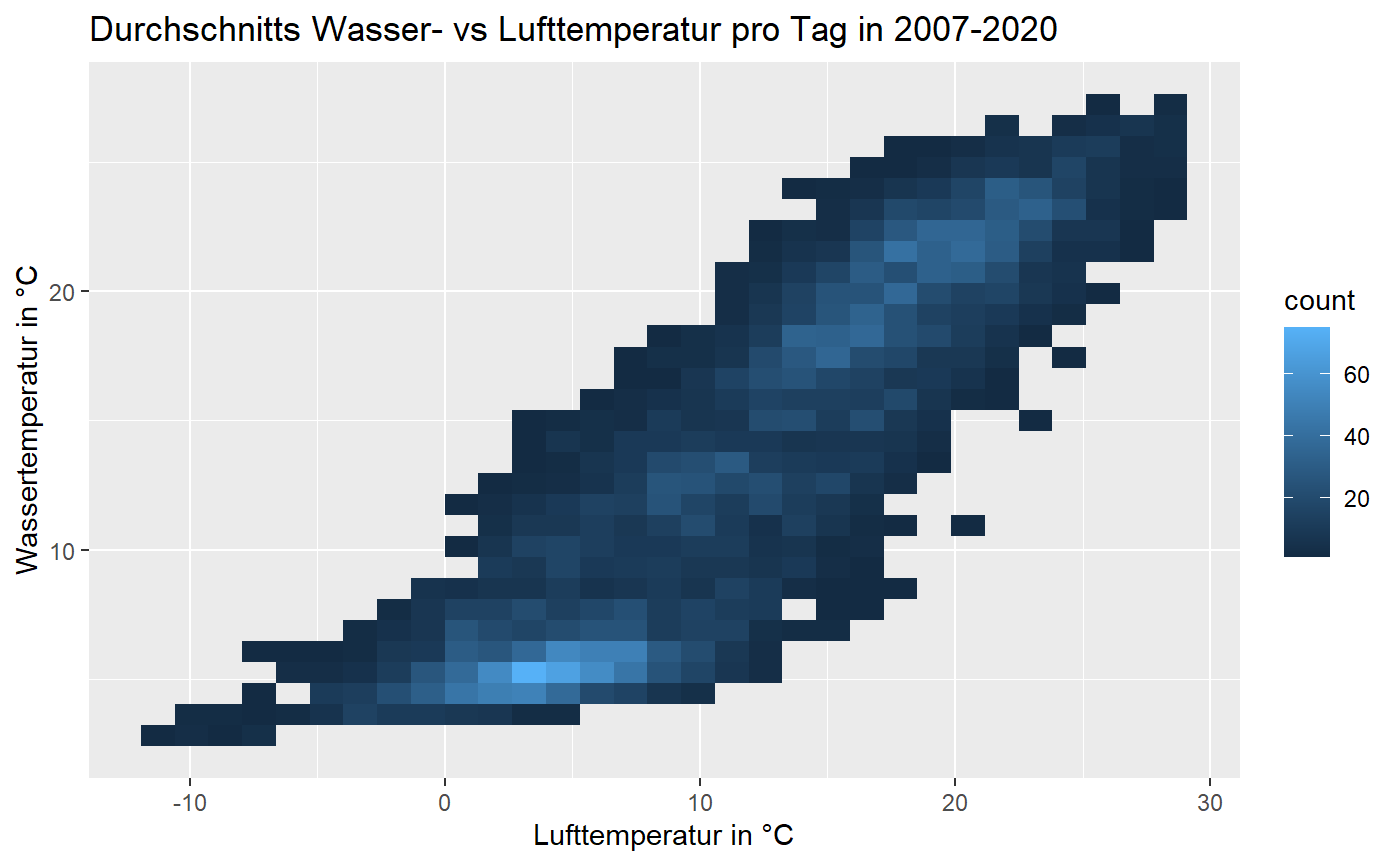
Chart, histogram

Description automatically generated

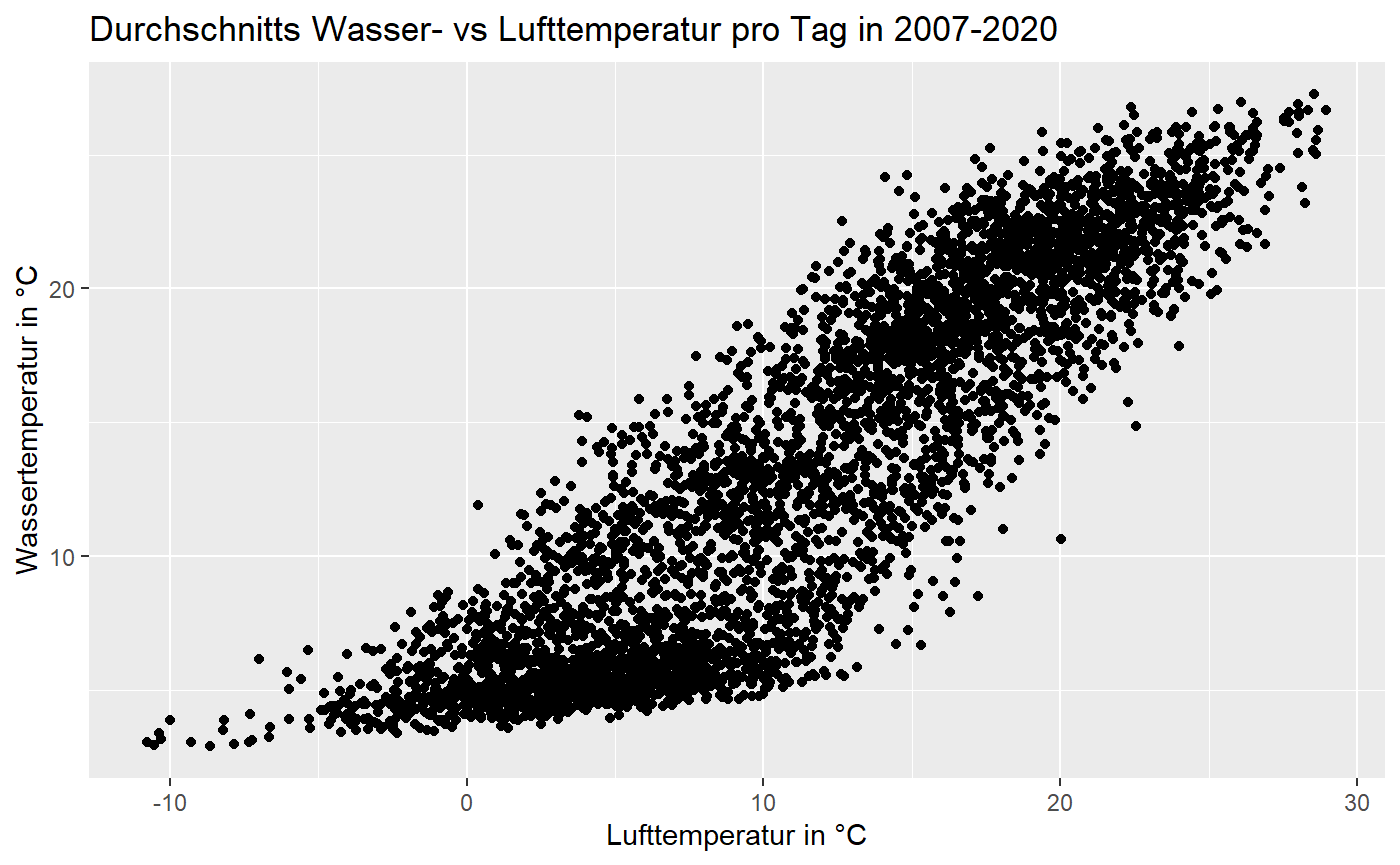
## Abbildung 4: Boxplot über die Verteilung des totalen Niederschlags pro Tag in Mythenquai



## Abbildung 5: Heatmap: Wasser- vs Luft-temperatur in 2007-2020



## Abbildung 6: Punktewolke: Wasser- vs Luft-temperatur in 2007-2020



## Abbildung 7: Punktewolke: Wasser- vs Lufttemperatur in 2018

Chart, scatter chart

Description automatically generated

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



## Abbildung 9: Windrichtung und Geschwindigkeit in Mythenquai im 2018

Chart, radar chart

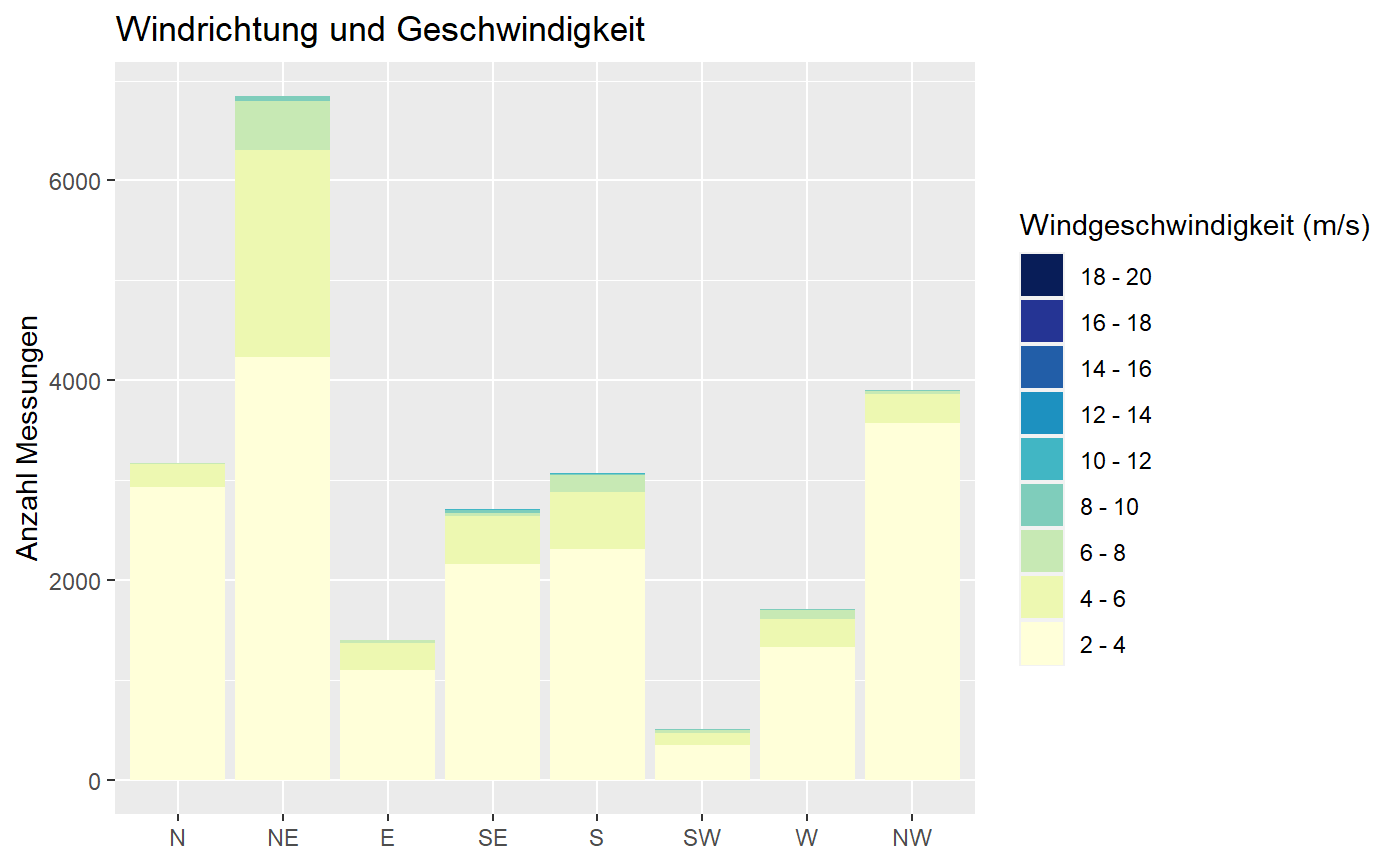
Description automatically generated

## Abbildung 10: Hierarchie der Grammar of Graphics

A picture containing chart

Description automatically generated

## Abbildung 11: Windrose im kartesischen Koordinatensystem



## Abbildung 12: Windrichtung mit Windgeschwindigkeit als Facetten

Chart, bar chart

Description automatically generated