Grundlagen Datenvisualisierungen

Bericht

Student: Florin Barbisch

Studiengang: FHNW, Data Science

Fachexpertin: Arzu Cöltekin

Datum: 3.17.2022

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 auseinander setzen möchte.

Alle Diagramm 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 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 : 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

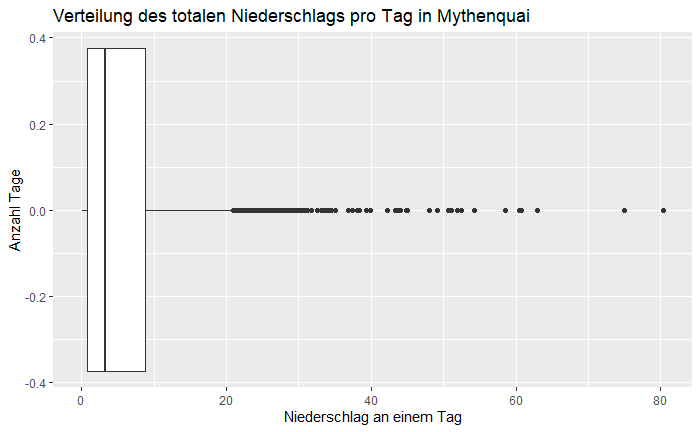
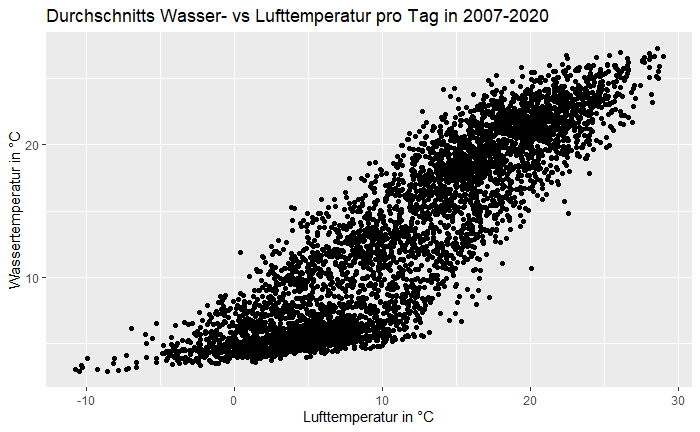
Eine Alternative zum Histogramm ist der Boxplot. Dort werden in einer mit einer Box die mittleren 50% der Daten markiert. Der Strich in der Box zeigt den median an. Die Antennen in beide Richtungen sind 1.5 mal so lang wie der Interquartilabstand. Punkte welche ausserhalb dieser Box und der Antenne liegen werden einzeln markiert (Yi, n.d.-a).

Abbildung : 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.

## Punktewolken/Heatmap

Chart, histogram

Description automatically generated

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

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

Abbildung : 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.

# Quellenverzeichnis

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

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

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.

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

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

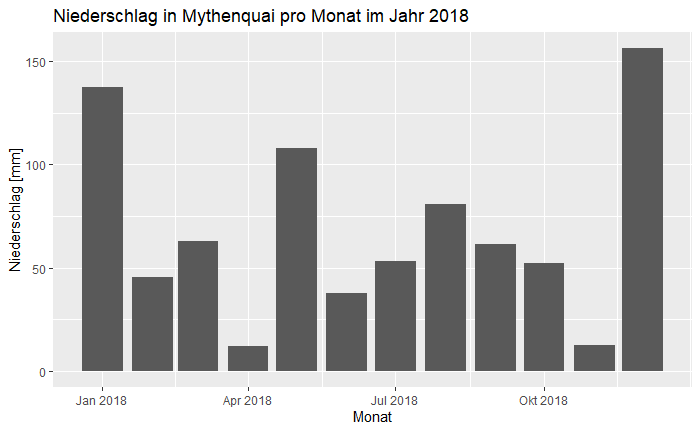
Abbildung 1: Balkendiagramm über den Niederschlag in Mythenquai pro Monat im Jahr 2018

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

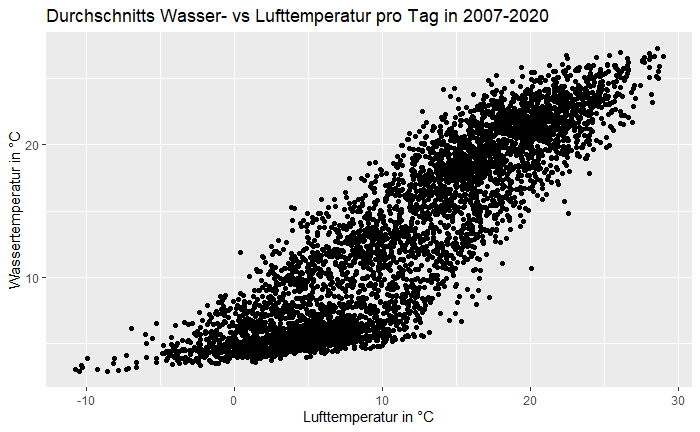


Abbildung 5: Heatmap: Wasser- vs Lufttemperatur in 2007-202

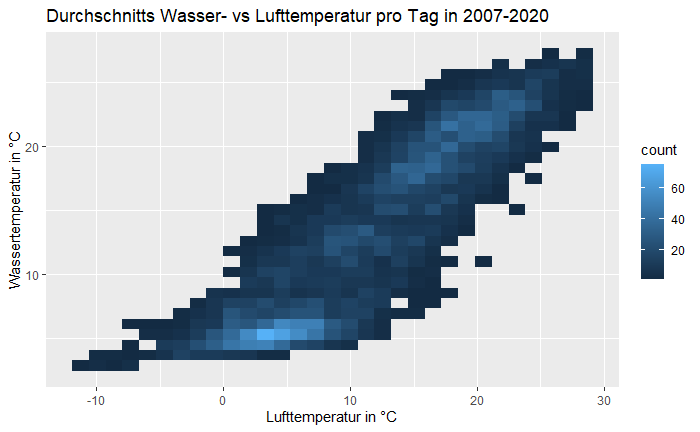


Abbildung 6: Punktewolke: Wasser- vs Lufttemperatur in 2018

