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. Der Code befindet sich auf einem GitHub-Repository (Barbisch, 2022).

Balkendiagramm

Bei 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.

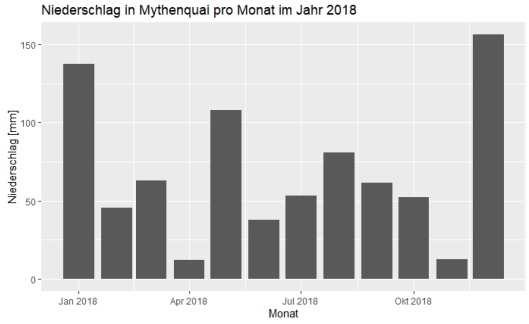

Abbildung 1: Balkendiagramm über den Niederschlag in Mythenquai pro Monat im Jahr 2018

Liniendiagramm

Ein 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). Deshalb macht es keinen Sinn kategorielle Daten auf der x-Achse darzustellen.

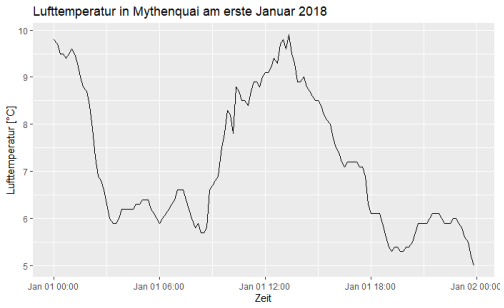

Abbildung 2: Liniendiagramm über den Verlauf der Lufttemperatur in Mythenquai am ersten Januar 2018

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. In der Abbildung 2 wird der Temperaturverlauf über die Zeit in einem Liniendiagramm dargestellt. Auf der x-Achse ist alle 6 Stunden eine Beschriftung der Achse. Eine fallende Linie zeigt hier dem Betrachter klar, dass es kälter wurde. Eine gerade Linie zeigt, dass die Temperatur konstant blieb.

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.

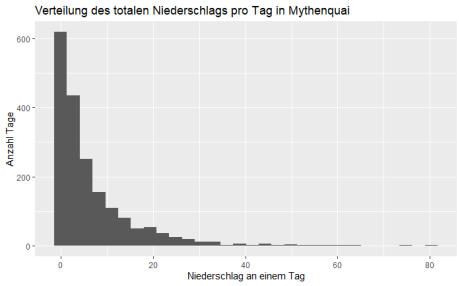

Abbildung 3: Histogramm über die Verteilung des totalen Niederschlags pro Tag in Mythenquai

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 $numbins = \lceil \log_2 \mathrm{n} \rceil + 1$ entwickelt um die Anzahl Balken zu ermitteln.

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.

Punktewolken/Heatmap

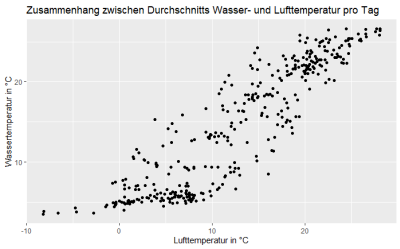

Abbildung 6: Punktewolke: Wasser- vs Lufttemperatur in 2018

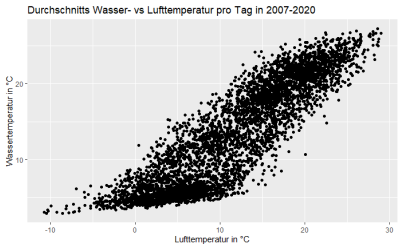

Abbildung 4: Punktewolke: Wasser- vs Lufttemperatur in 2007-2020

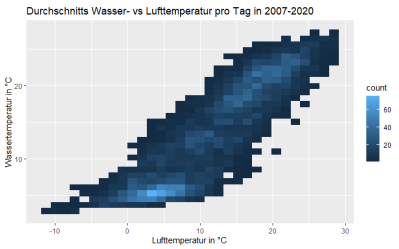

Abbildung 5: Heatmap: Wasser- vs Lufttemperatur in 2007-2020

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.

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).

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.