Universität Stuttgart

Projekt Data Science
Analyse von Mobilitätsdaten

Abschlussvortrag

Wintersemester 2023/24

Gruppe 02 – Ozan Tastekin

Warum eine Analyse von Bahndaten?
Geschäftsbericht der SSB

341 Mio.

Fahrten im VVS gesamt

149 Mio.

Fahrten nur mit SSB-Verkehrsmitteln

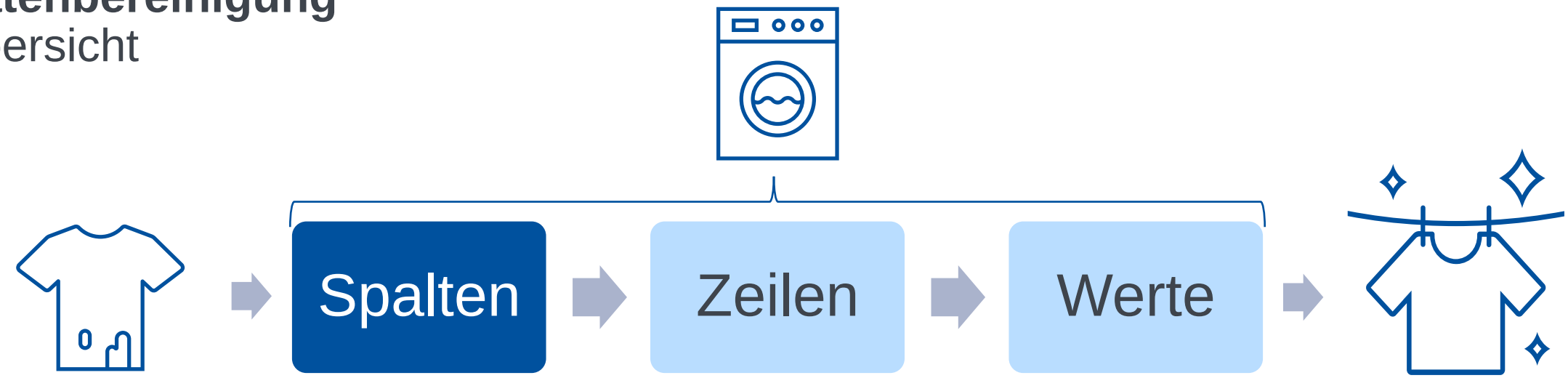
151 Mio. €

netto Ticketerlöse

Quelle: https://www.ssb-ag.de/unternehmen/informationen-fakten/geschaeftsberichte/geschaeftsbericht-2022/

Datenbereinigung

Datenbereinigung
Übersicht



Spalten entfernt, falls…

- Spalte keine Information gibt

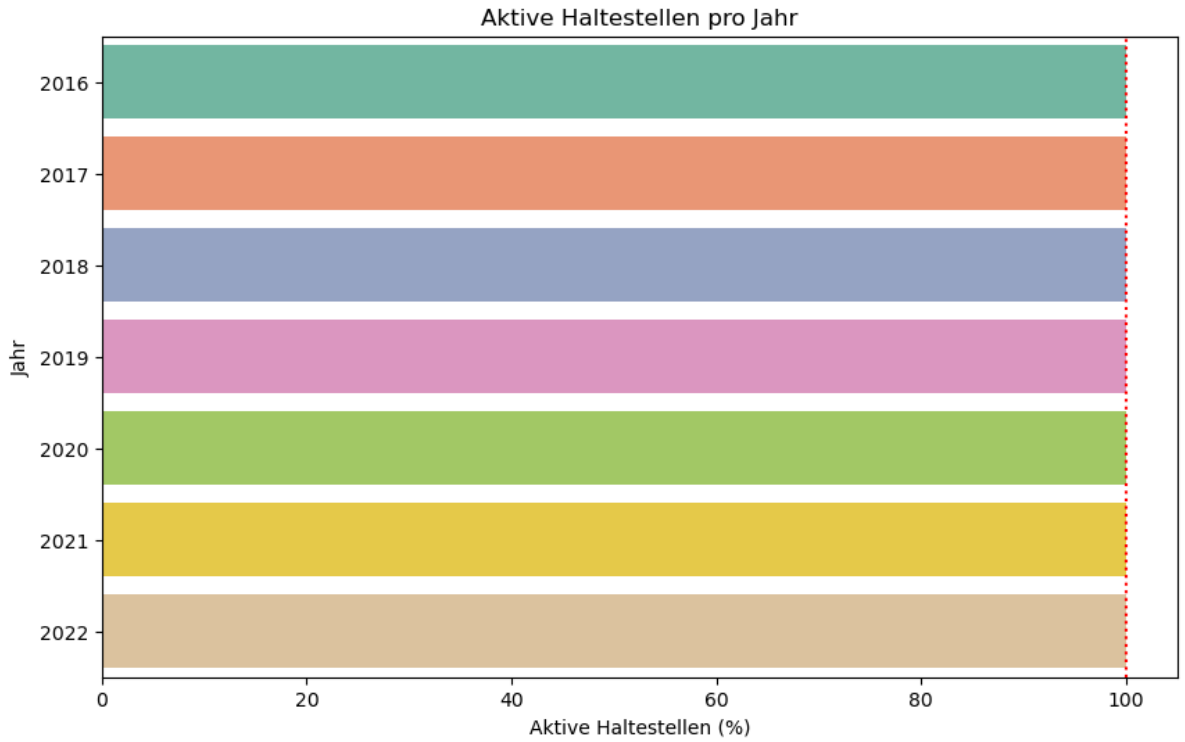
oder

- Spalte an anderer Stelle dieselbe Information hat

Datenbereinigung – Spalten
Keine Information

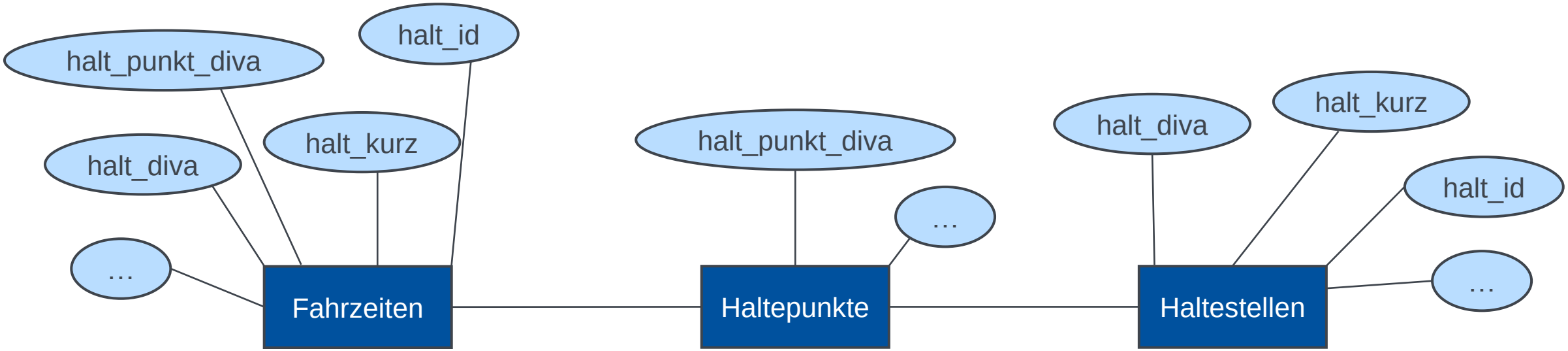
Spalten

Haltestellen



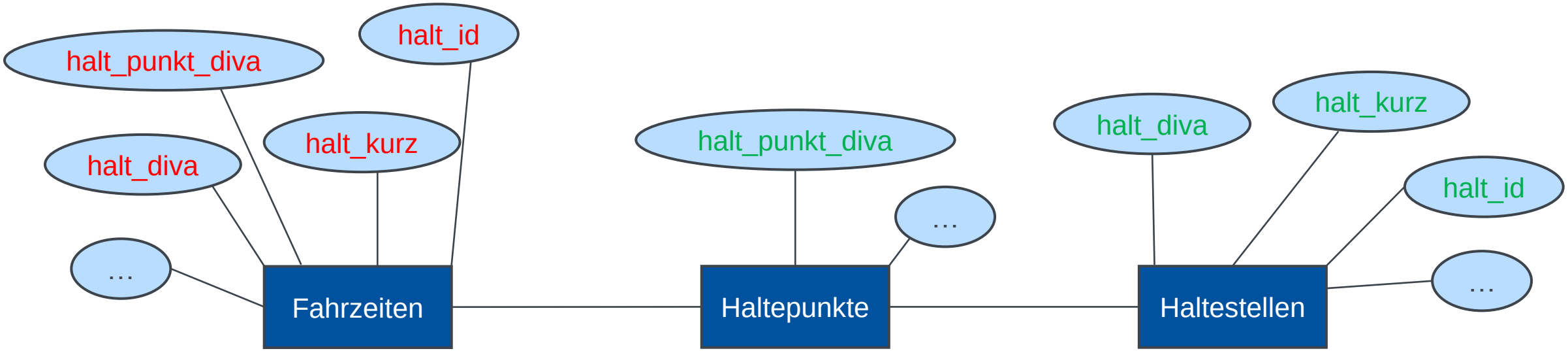
Datenbereinigung – Spalten

An anderer Stelle dieselbe Information

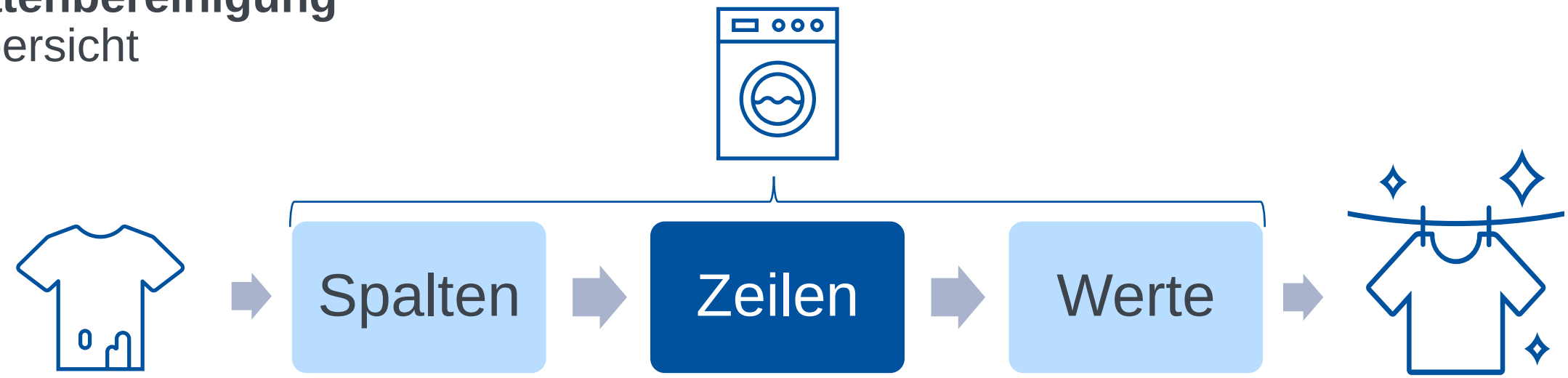


Datenbereinigung – Spalten

An anderer Stelle dieselbe Information



Datenbereinigung
Übersicht



Zeilen entfernt, falls…

- Zeilen dupliziert vorliegen

Datenbereinigung – Zeilen
Duplizierte Zeilen

Zeilen

Nur in Fahrzeiten gab es duplizierte Zeilen

12130 (≪ 0,001%) duplizierte Zeilen entfernt

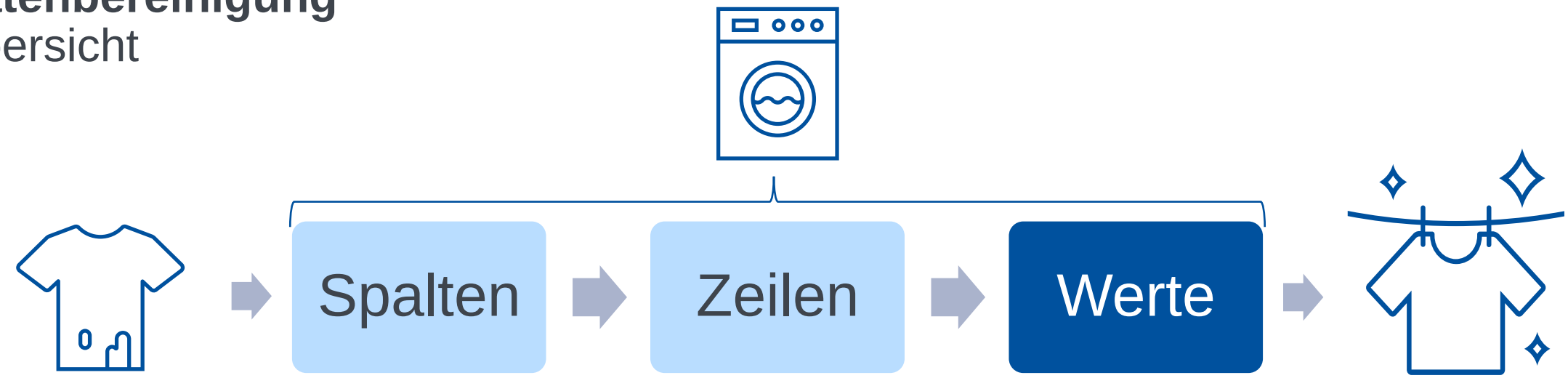
	linie	richtung	betriebsdatum	fahrzeug	kurs	seq_von	halt_diva_von	halt_punkt_diva_von	halt_kurz_von1	datum_von	soll_an_von
0	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	22140
1	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	22140
2	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	22140
3	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	22140
4	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	25680
5	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	25680

Zeilen

entfernt

	linie	richtung	betriebsdatum	fahrzeug	kurs	seq_von	halt_diva_von	halt_punkt_diva_von	halt_kurz_von1	datum_von	soll_an_von
0	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	22140
1	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	25680
2	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	29280
3	314	1	13.01.22	11443	2	2	2406.0	50.0	SOME	13.01.22	22200
4	314	1	13.01.22	11443	2	2	2406.0	50.0	SOME	13.01.22	25752
5	314	1	13.01.22	11443	2	2	2406.0	50.0	SOME	13.01.22	29352

Datenbereinigung
Übersicht



Fehlende Werte wurden…

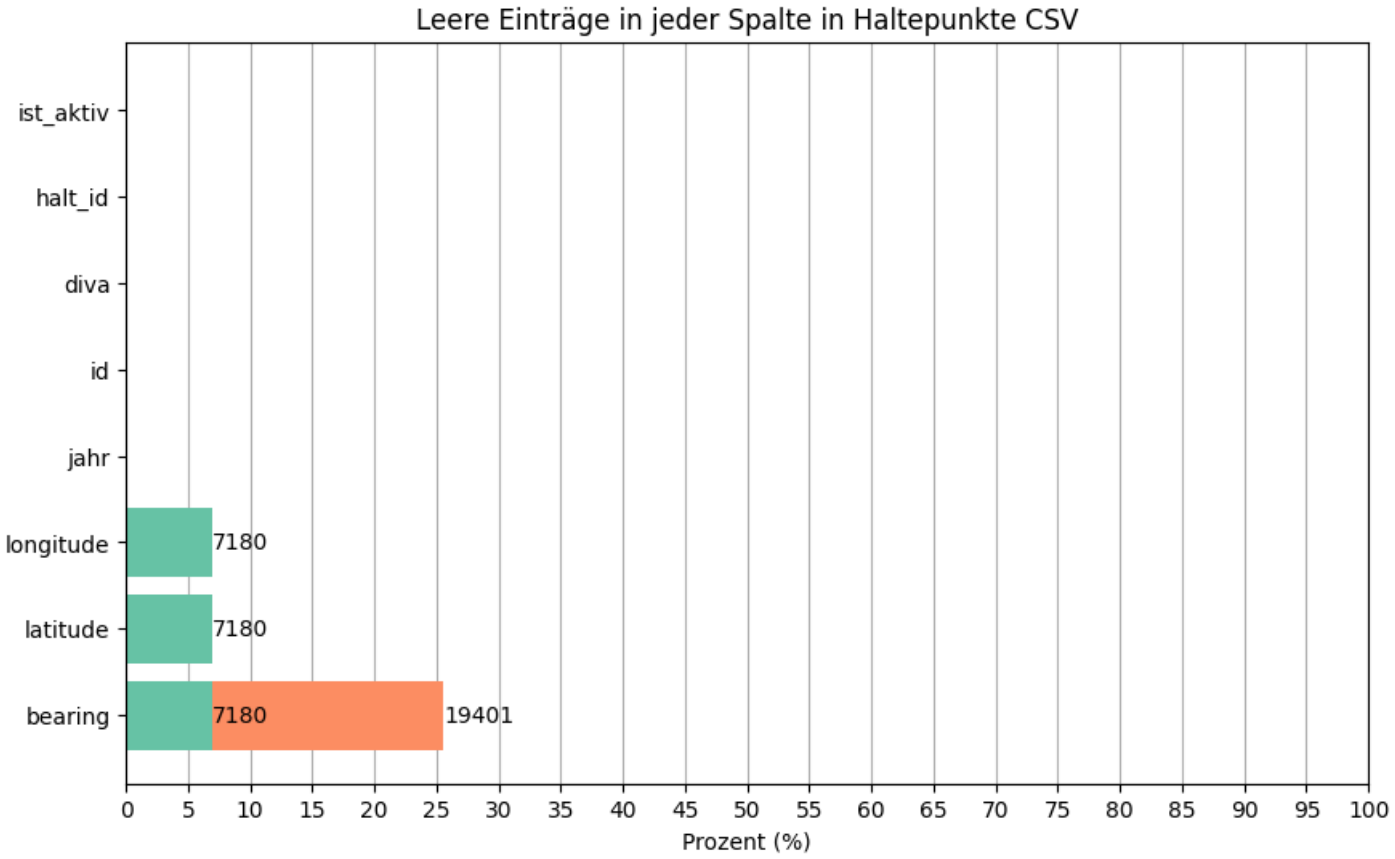
- sinnvoll aufgefüllt

Datenbereinigung – Werte

Werte sinnvoll aufgefüllt

Werte

Haltepunkte



Hypothesen

Langer Tag an der Uni. S-Bahnen fahren nicht! Der Bus an der Universität Haltestelle wird verspätet sein... Warum?

Hypothese 1

Im Jahr 2022 gibt es einen positiven linearen Zusammenhang zwischen ein- und aussteigenden Fahrgästen und aufgebauten Verspätungen an der Hardbrücke Haltestelle bei den Linien 33, 72, 83 und 8.

H1 – Verspätung durch viele Passagiere
Benötigte Daten

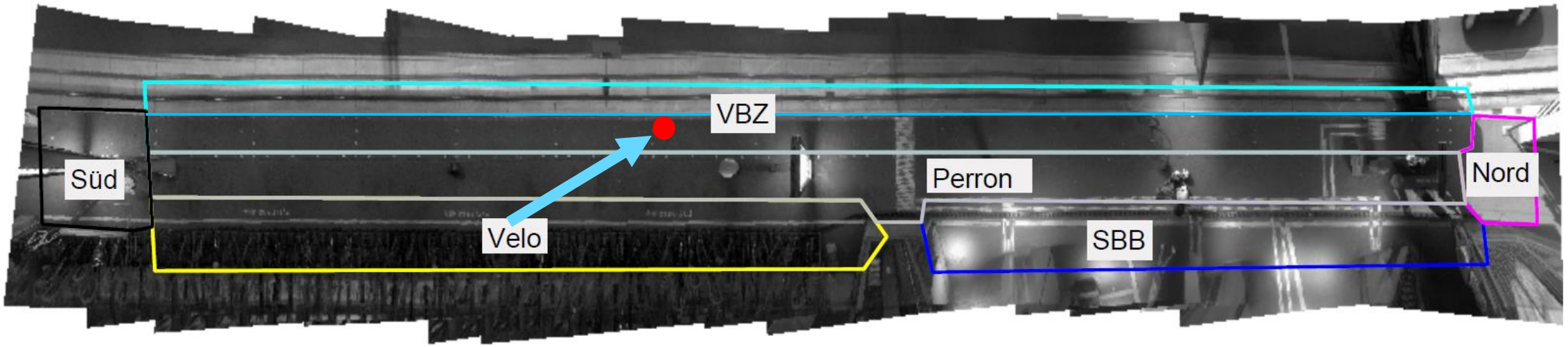
- Fahrzeiten Soll und Ist-Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022
- Passagierfrequenzen an der Hardbrücke Haltestelle in Zürich[2]
  - Ein- und Aussteiger an 2 Gleisen (Hin- und Rückrichtung)
    - Buslinien 33, 72, 83 und Tramlinie 8 sind davon betroffen
  - Gleise sind auf separaten Bahnsteigen
  - Alle **5 Minuten** gezählt
  - 2020-2023

[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

[2] https://data.stadt-zuerich.ch/dataset/vbz_frequenzen_hardbruecke

H1 – Verspätung durch viele Passagiere
Benötigte Daten – Passagierfrequenzen

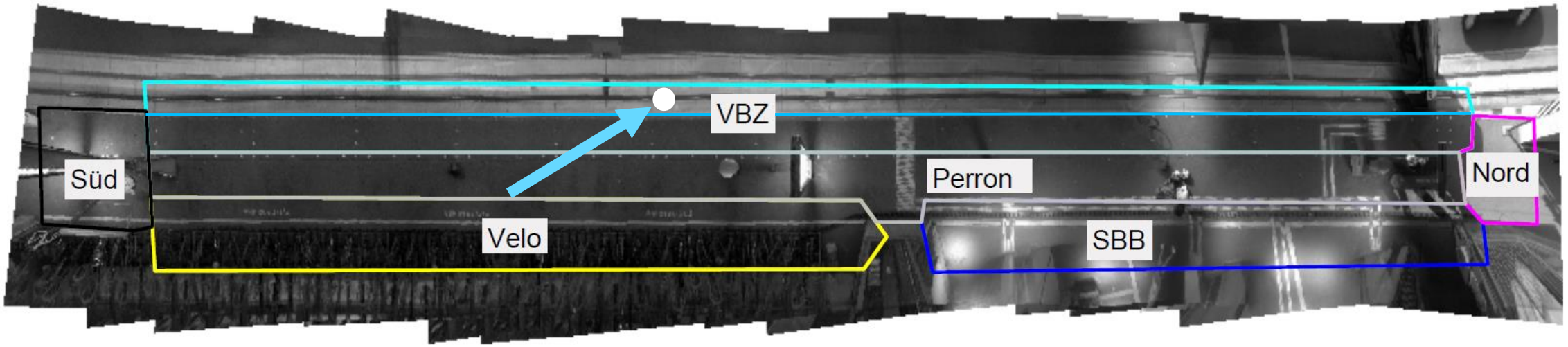
Gleise & Sensoren des Datensatzes



Quelle: https://data.stadt-zuerich.ch/dataset/vbz_frequenzen_hardbruecke

H1 – Verspätung durch viele Passagiere
Benötigte Daten – Passagierfrequenzen

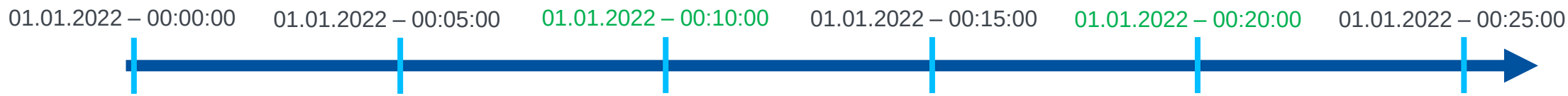
Gleise & Sensoren des Datensatzes



Quelle: https://data.stadt-zuerich.ch/dataset/vbz_frequenzen_hardbruecke

H1 – Verspätung durch viele Passagiere
Vorgehen

1. Passagierfrequenz Daten bereinigt
   - 11% der Zeitstempel (Zeilen) fehlen

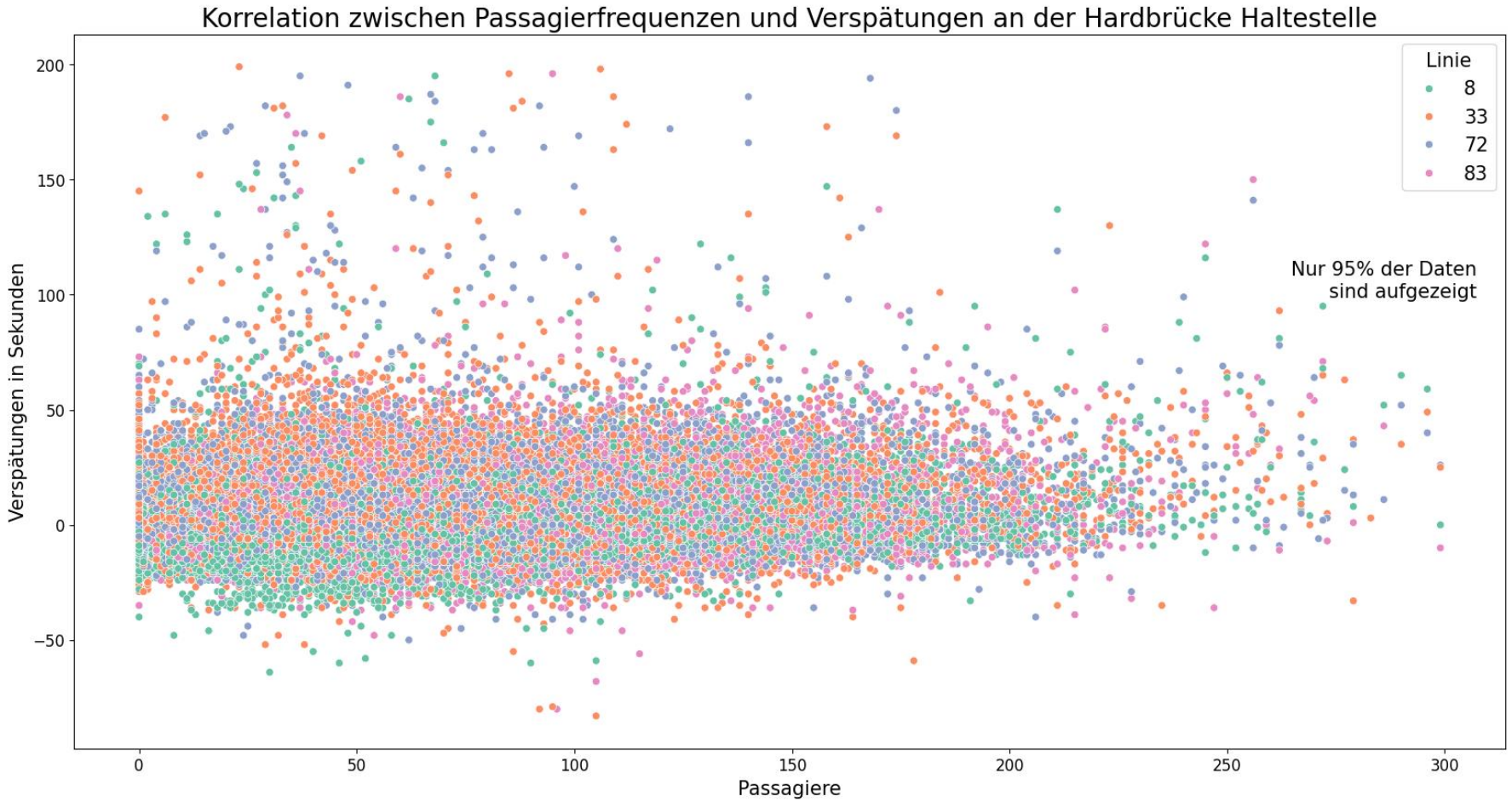


Einsteiger	2	0	0	1	0	3
Aussteiger	0	4	0	1	0	5

H1 – Verspätung durch viele Passagiere
Vorgehen

1. Passagierfrequenz Daten bereinigt
   - 11% der Zeitstempel (Zeilen) fehlen ⟹ Zeilen aufgefüllt
2. Daten transformiert
   - Zeitstempel auf das gleiche Format gebracht
   - Haltepunkte von Fahrzeiten zu Gleisen von Passagierfrequenzen zugeordnet
   - Verspätungen ausgerechnet & auf 5 Minuten zusammengerechnet
   - Passagierfrequenzen und Verspätungen verbunden
3. Analyse durchgeführt
   - Spearman-Rangkorrelation durchgeführt
     - Geeignete Korrelationsanalyse, wenn Daten nicht normalverteilt sind (hier der Fall)
     - Ausgabe: Korrelationskoeffizient $-1 \leq r \leq 1$
       - $0{,}5 \leq r < 0{,}7$: hohe positive Korrelation
       - $0{,}7 \leq r \leq 1$: sehr hohe positive Korrelation
     - **Annahmekriterium:** $r \geq 0{,}5$

H1 – Verspätung durch viele Passagiere
Ergebnisse



r = 0,1529 ⟹ Hypothese wird abgelehnt

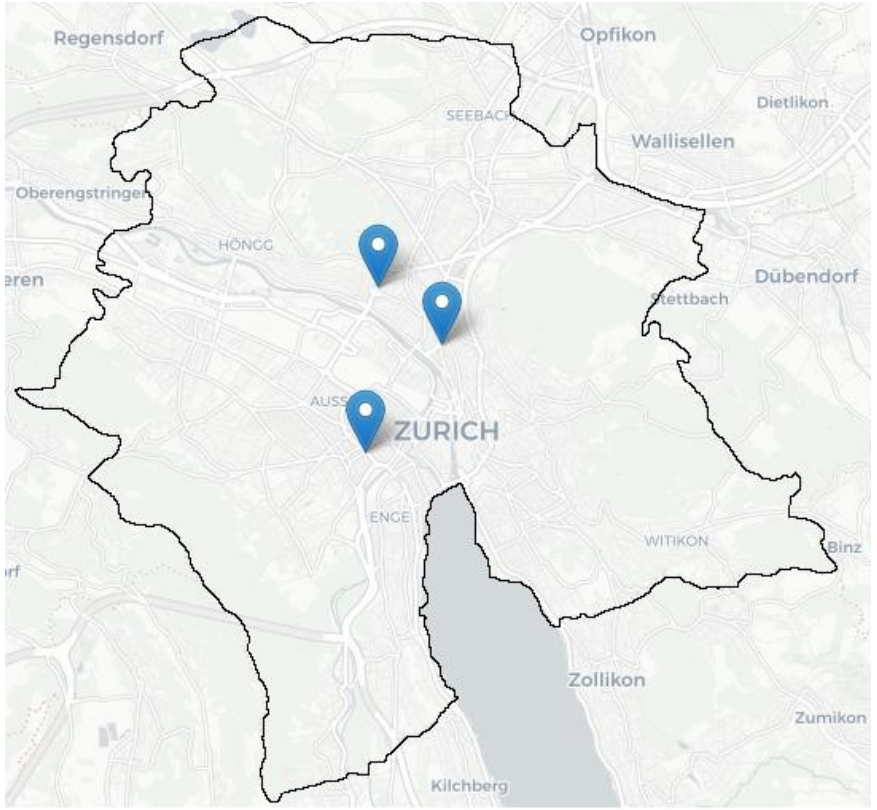
Bei Regen verdoppelt sich der Bremsweg, merk dir das für die Theorie Prüfung!

Hypothese 2

In der Mehrheit der Jahre von 2016 bis 2022 verzeichneten Busse in Stunden mit mehr als 30 Minuten Regen überdurchschnittlich hohe aufgebaute Verspätungen.

H2 – Verspätung durch Regen
Benötigte Daten

- Fahrzeiten Soll und Ist-Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022
- Stündliche Wetterdaten in Zürich[2]
  - 3 Standorte der Messung in Zürich
  - Temperatur, **Regendauer**, Luftdruck, Luftfeuchtigkeit, etc.
  - 2000-2023



[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

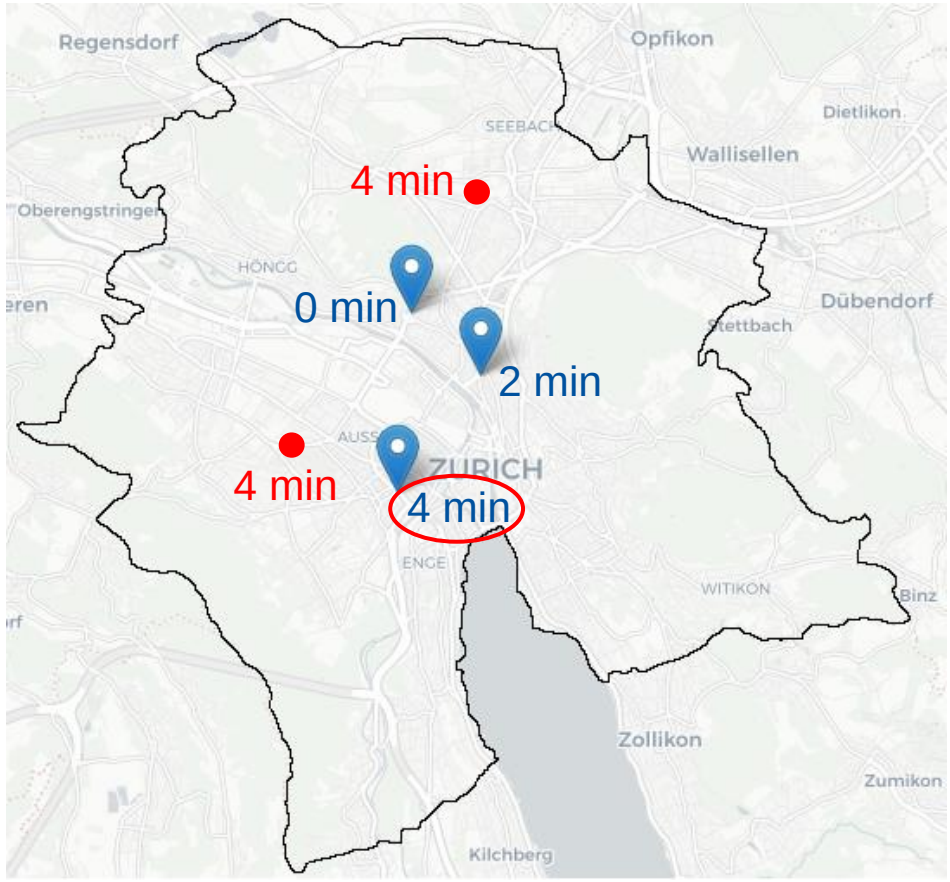
[2] https://data.stadt-zuerich.ch/dataset/ugz_meteodaten_stundenmittelwerte

H2 – Verspätung durch Regen
Vorgehen

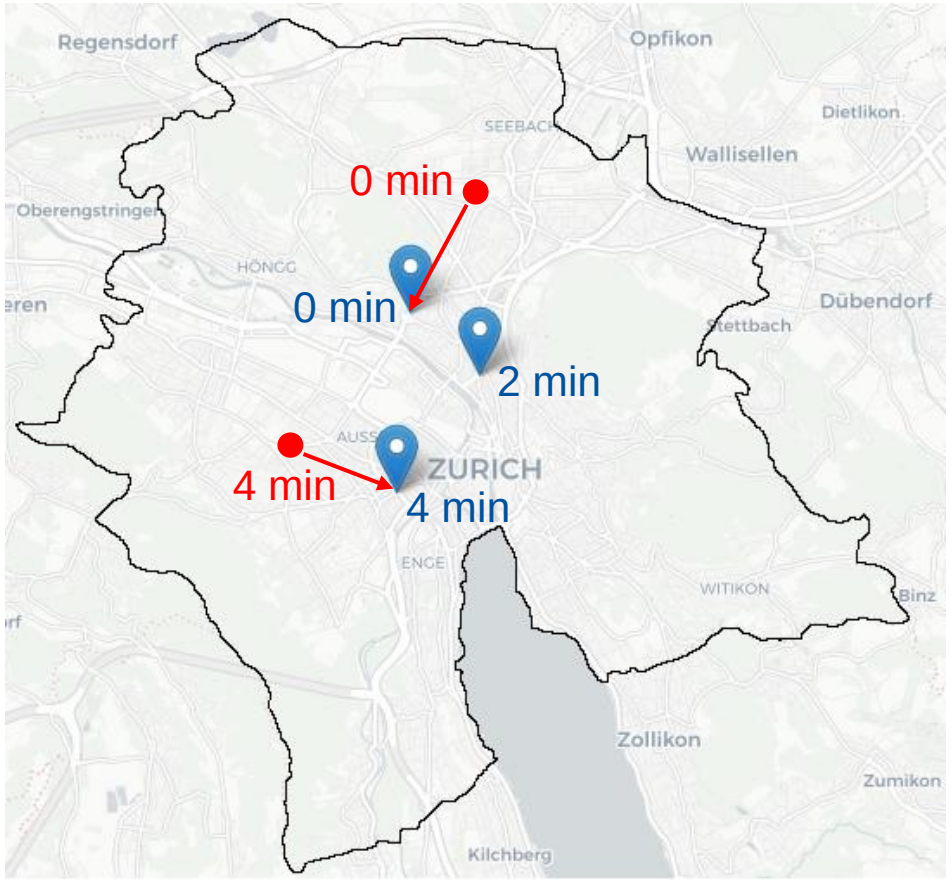
1. Wetter Daten bereinigt
2. Daten transformiert
   - Zeitstempel auf das gleiche Format gebracht
   - Verspätungen ausgerechnet & auf 1 Stunde zusammengerechnet
   - Regendauer und Verspätungen verbunden

H2 – Verspätung durch Regen
Vorgehen

Version 1: Maximale Regendauer



Version 2: Nächste Wetterstation

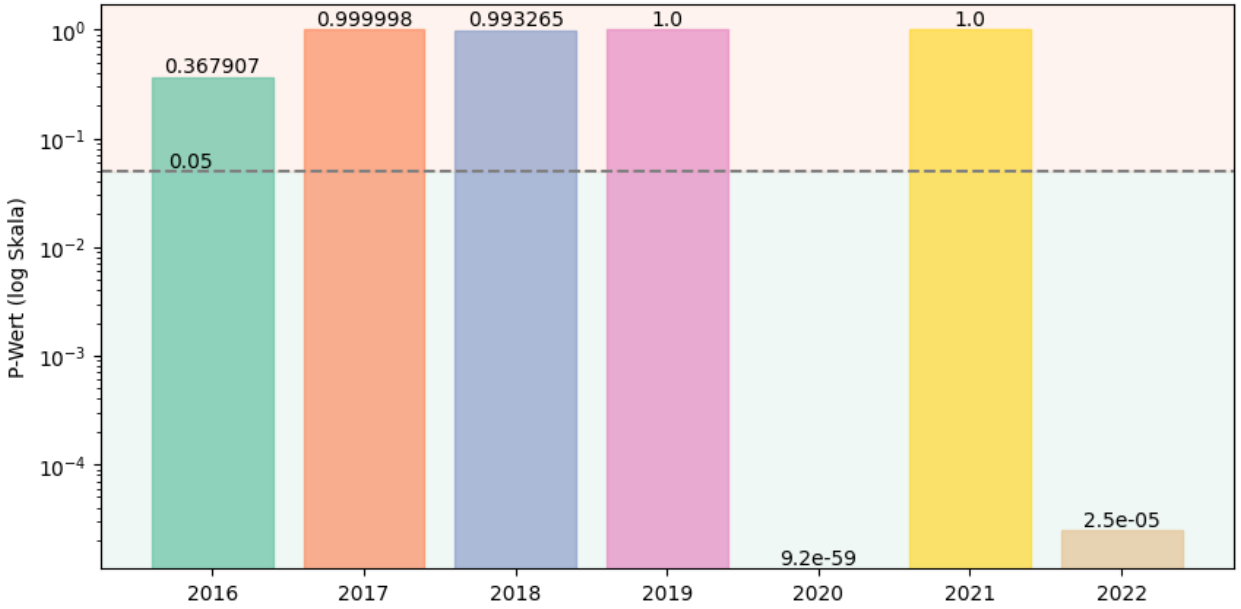


H2 – Verspätung durch Regen
Vorgehen

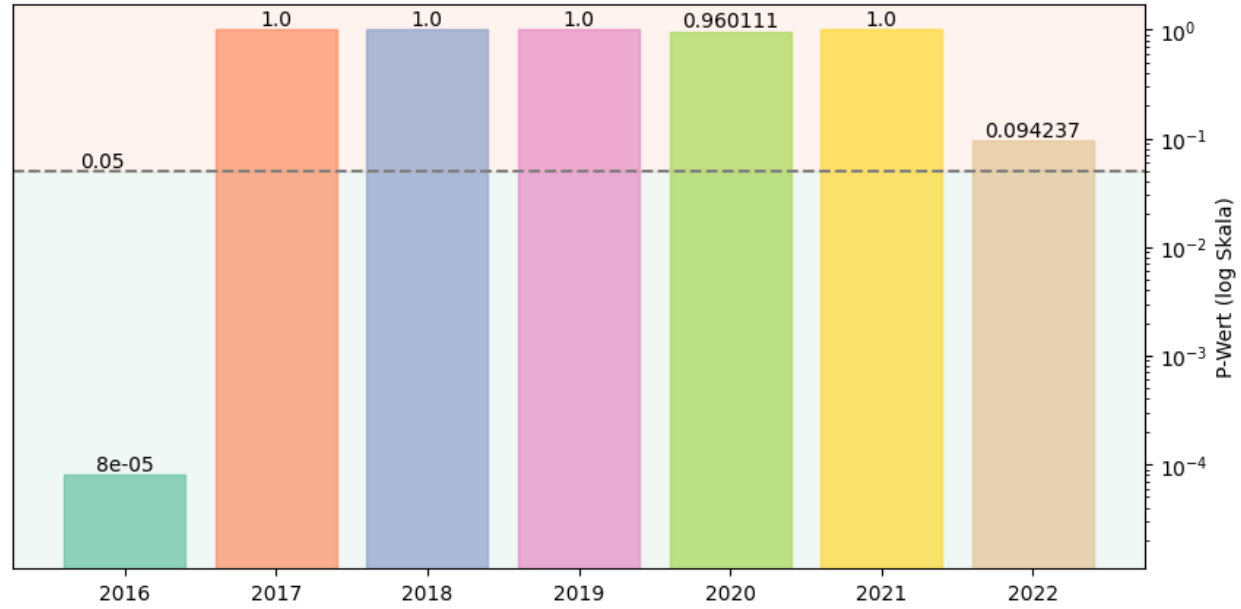
1. Wetter Daten bereinigt
2. Daten transformiert
   - Zeitstempel auf das gleiche Format gebracht
   - Verspätungen ausgerechnet & auf 1 Stunde zusammengerechnet
   - Regendauer und Verspätungen verbunden
   - In 2 Kategorien eingeordnet: > 30 Minuten Regen und ≤ 30 Minuten Regen
3. Analyse durchgeführt
   - Mann-Whitney U Test durchgeführt
     - Geeigneter statistischer Signifikanztest, wenn Daten nicht normalverteilt sind (hier der Fall)
     - Ausgabe: Signifikanzniveau $0 \leq p \leq 1$
       - Je größer p, desto wahrscheinlicher, dass die beobachteten Ergebnisse ein Zufall sind
     - **Annahmekriterium:** $p \leq 0{,}05$ bei mindestens 4 der 7 Jahre (Mehrheit)

H2 – Verspätung durch Regen
Ergebnisse

Version 1: Maximale Regendauer



Version 2: Nächste Wetterstation



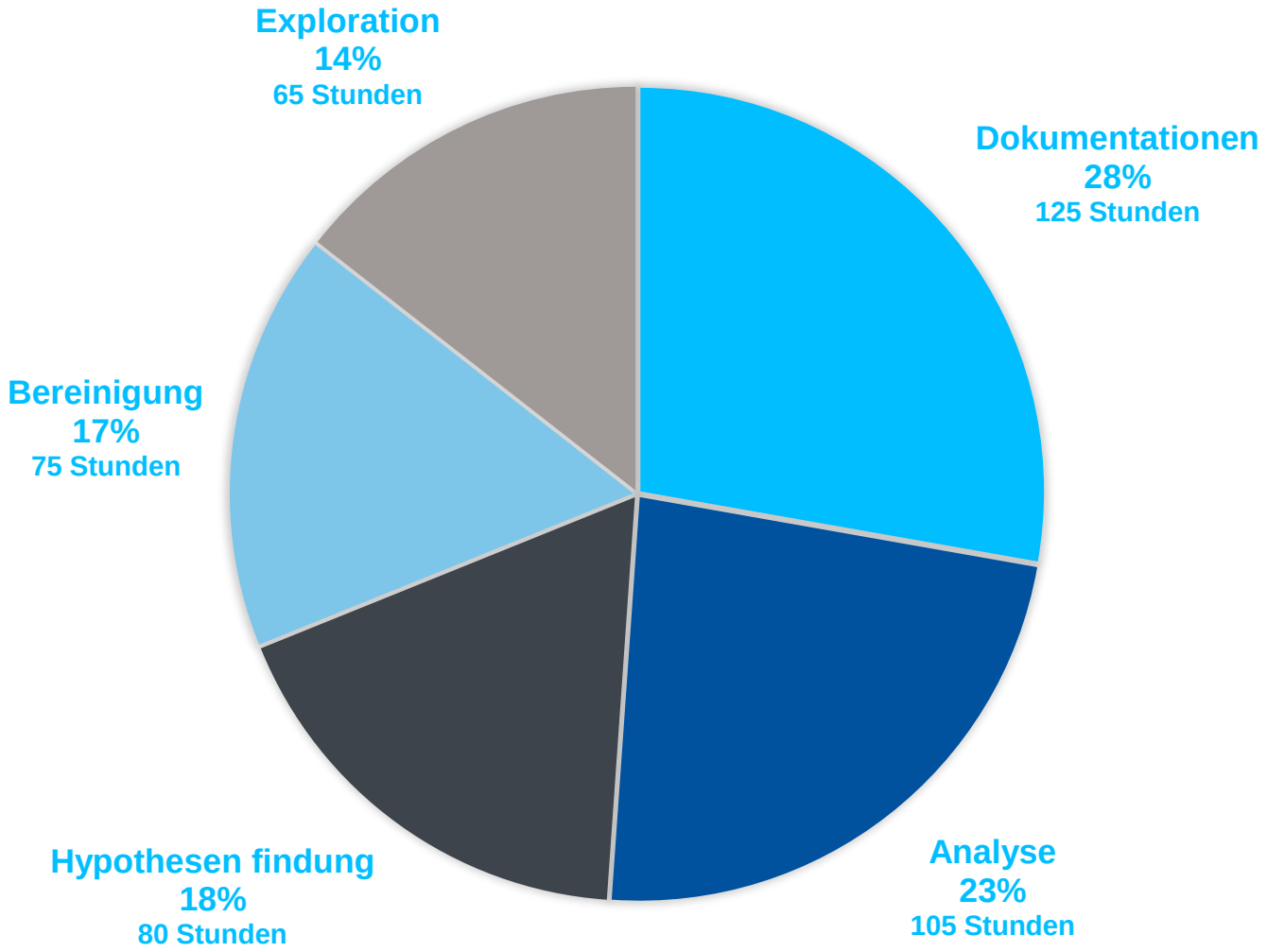
Bei beiden Versionen:

Annahmekriterium trifft bei weniger als 4 Jahren zu

⟹ Hypothese wird abgelehnt

Zeitaufwand

Zeitaufwand
Gesamt



Gesamtstunden:
450 Stunden