



Universität Stuttgart

Projekt Data Science
Analyse von Mobilitätsdaten

Datenbereinigung

Wintersemester 2023/24
Gruppe 02 – Ozan Tastekin & Tony Klasan

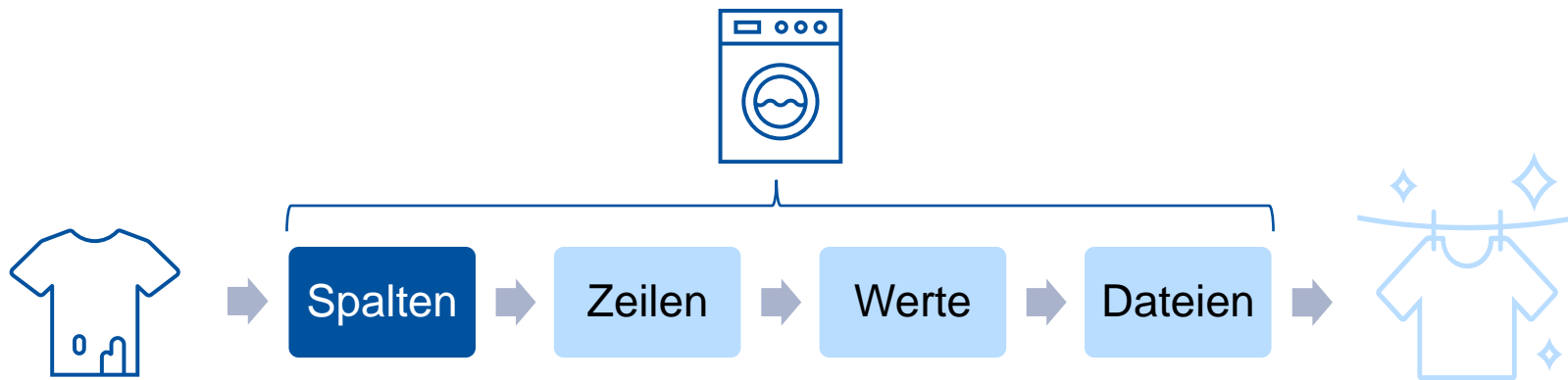
Letztes Mal bei...

Verständnis der Daten

- Eine Haltestelle kann mehrere Haltepunkte haben
- Was ist ein Haltepunkt?
 - Bahn- oder Bussteige, wo Passagiere ein- und aussteigen können

Datenbereinigung

Übersicht



Datenbereinigung – Spalten

Umbenennen & Entfernen



Spalten



Fahrzeiten		
linie	soll_ab_von	fahrtweg_id
richtung	ist_ab_von	fw_no
betriebsdatum	seq_nach	fw_typ
fahrzeug	halt_diva_nach	fw_kurz
kurs	halt_punkt_diva_nach	fw_lang
seq_von	halt_kurz_nach1	umlauf_von
halt_diva_von	datum_nach	halt_id_von
halt_punkt_diva_von	soll_an_nach	halt_id_nach
halt_kurz_von1	ist_an_nach1	halt_punkt_id_von
datum_von	soll_ab_nach	halt_punkt_id_nach
soll_an_von	ist_ab_nach	
ist_an_von	fahrt_id	

Haltepunkte
halt_punkt_id
halt_punkt_diva
halt_id
GPS_Latitude
GPS_Longitude
GPS_Bearing
halt_punkt_ist_aktiv

Haltestellen
halt_id
halt_diva
halt_kurz
halt_lang
halt_ist_aktiv

Datenbereinigung – Spalten

Umbenennen & Entfernen



Spalten



Fahrzeiten		
linie	soll_ab_von	fahrtweg_id
richtung	ist_ab_von	fw_no
betriebsdatum	seq_nach	fw_typ
fahrzeug	halt_diva_nach	fw_kurz
kurs	halt_punkt_diva_nach	fw_lang
seq_von	halt_kurz_nach1	umlauf_von
halt_diva_von	datum_nach	halt_id_von
halt_punkt_diva_von	soll_an_nach	halt_id_nach
halt_kurz_von1	ist_an_nach1	halt_punkt_id_von
datum_von	soll_ab_nach	halt_punkt_id_nach
soll_an_von	ist_ab_nach	
ist_an_von	fahrt_id	

Haltepunkte
halt_punkt_id
halt_punkt_diva
halt_id
GPS_Latitude
GPS_Longitude
GPS_Bearing
halt_punkt_ist_aktiv

Haltestellen
halt_id
halt_diva
halt_kurz
halt_lang
halt_ist_aktiv

Datenbereinigung – Spalten


Umbenennen & Entfernen



Fahrzeiten

rs	seq_von	halt_diva_von	halt_punkt_diva_von	halt_kurz_von1	datum_von	soll_an_von
1	27	1456	0	KRES	04.01.22	2490
1	26	1845	0	OPER	04.01.22	2480
13	27	1456	0	KRES	04.01.22	2400
13	26	1845	0	OPER	04.01.22	2390
13	9	1456	0	KRES	04.01.22	1820
13	8	1845	0	OPER	04.01.22	1810
2	27	1456	0	KRES	04.01.22	2530


Spalte
entfernt



rs	seq_von	datum_von	soll_an_von
1	27	04.01.22	2490
1	26	04.01.22	2480
13	27	04.01.22	2400
13	26	04.01.22	2390
13	9	04.01.22	1820
13	8	04.01.22	1810
2	27	04.01.22	2530

on	seq_nach	halt_diva_nach	halt_punkt_diva_nach	halt_kurz_nach1	datum_nach	soll_an_nach
28	28	832	0	FELD	04.01.22	2490
74	27	1456	0	KRES	04.01.22	2490
35	28	832	0	FELD	04.01.22	2400
35	27	1456	0	KRES	04.01.22	2400
32	10	832	0	FELD	04.01.22	1830
76	9	1456	0	KRES	04.01.22	1830

Spalte
entfernt



on	seq_nach	datum_nach	soll_an_nach
28	28	04.01.22	2490
74	27	04.01.22	2490
35	28	04.01.22	2400
35	27	04.01.22	2400
32	10	04.01.22	1830
76	9	04.01.22	1830

Datenbereinigung – Spalten

Umbenennen & Entfernen



Fahrzeiten

u_lang	umlauf_von	halt_id_von	halt_id_nach	halt_punkt_id_von	halt_punkt_id_nach
- BTIE	262627	2612	2689	28468	10560
- BTIE	262627	2104	2612	10538	28468
- BTIE	262671	2612	2689	28468	10560
- BTIE	262671	2104	2612	10538	28468
isfahrt	262671	2612	2689	28468	10560
isfahrt	262671	2104	2612	10538	28468

Spalte
entfernt

u_lang	umlauf_von	halt_punkt_id_von	halt_punkt_id_nach
- BTIE	262627	28468	10560
- BTIE	262627	10538	28468
- BTIE	262671	28468	10560
- BTIE	262671	10538	28468
isfahrt	262671	28468	10560
isfahrt	262671	10538	28468

8 Spalten weniger

34 Spalten → 26 Spalten

Datenbereinigung – Spalten

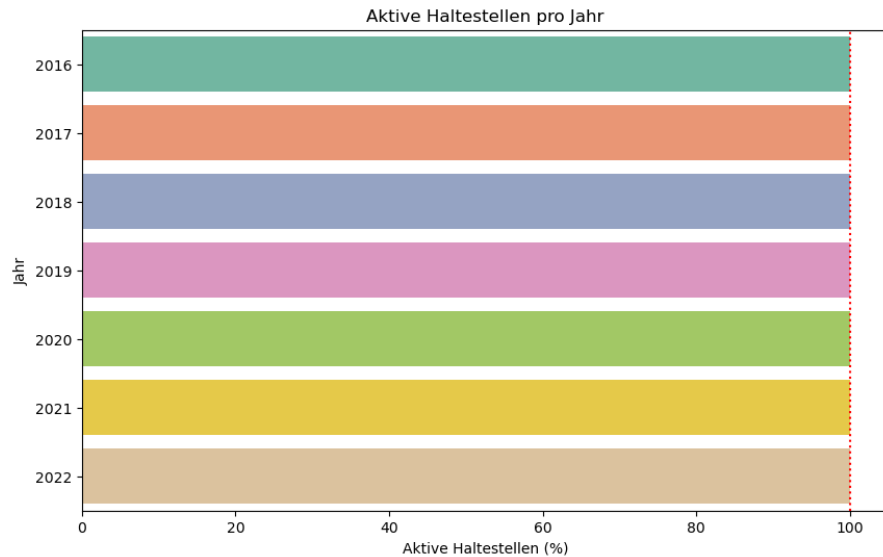
Umbenennen & Entfernen



Spalten



Haltestellen



Datenbereinigung – Spalten

Umbenennen & Entfernen



Haltestellen

	halt_id	halt_diva	halt_kurz	halt_lang	halt_ist_aktiv
0	143	2570	BirWSL	Birmensdorf ZH, Sternen/WSL	True
1	309	3356	WalBir	Waldegg, Birmensdorferstrasse	True
2	373	6232	FRAF07	Zürich Flughafen, Fracht	True
3	539	2655	TBAH01	Thalwil, Bahnhof	True
4	588	3027	FLUG07	Zürich Flughafen, Bahnhof	True
5	623	2989	TZEN01	Thalwil, Zentrum	True
6	701	1012	GOLB	Zürich, Goldbrunnenplatz	True

Spalte
entfernt

	id	diva	halt_kurz	halt_lang
0	143	2570	BirWSL	Birmensdorf ZH; Sternen/WSL
1	309	3356	WalBir	Waldegg; Birmensdorferstrasse
2	373	6232	FRAF07	Zürich Flughafen; Fracht
3	539	2655	TBAH01	Thalwil; Bahnhof
4	588	3027	FLUG07	Zürich Flughafen; Bahnhof
5	623	2989	TZEN01	Thalwil; Zentrum
6	701	1012	GOLB	Zürich; Goldbrunnenplatz

5 Spalten → 4 Spalten

Datenbereinigung – Spalten

Umbenennen & Entfernen



Haltepunkte

	halt_punkt_id	halt_punkt_diva	halt_id	GPS_Latitude	GPS_Longitude	GPS_Bearing	halt_punkt_ist_aktiv
0	303	51	143	47.360017	8.456337	85.0	False
1	304	50	143	47.360153	8.456180	270.0	False
2	686	50	309	47.368125	8.463072	212.0	False
3	687	51	309	47.368433	8.463819	19.0	False
4	823	51	373	47.452401	8.571871	208.0	False
5	824	50	373	47.452586	8.572158	29.0	False
6	825	1	373	47.452018	8.571423	92.0	False

Spalten
➡
umbenannt

	id	diva	halt_id	latitude	longitude	bearing	ist_aktiv
0	303	51	143	47.360017	8.456337	85.0	False
1	304	50	143	47.360153	8.456180	270.0	False
2	686	50	309	47.368125	8.463072	212.0	False
3	687	51	309	47.368433	8.463819	19.0	False
4	823	51	373	47.452401	8.571871	208.0	False
5	824	50	373	47.452586	8.572158	29.0	False
6	825	1	373	47.452018	8.571423	92.0	False

Datenbereinigung – Spalten

Umbenennen & Entfernen



Spalten



Passagierfrequenz

	code_codice	uic	bahnhof_gare_stazione	kt_ct_cantone	isb_gi	jahr_annee_anno
0	AAT	8503124.0	Aathal	ZH	SBB	2018.0
1	ABO	8502000.0	Aarburg-Oftringen	AG	SBB	2018.0
2	AIG	8501400.0	Aigle	VD	SBB	2018.0
3	AIG	8501400.0	Aigle	VD	SBB	2022.0
4	ALT	8506319.0	Altstätten SG	SG	SBB	2018.0
5	ALTD	8503211.0	Altendorf	SZ	SBB	2018.0
6	ALV	8509194.0	Alvaneid	GR	RhR	2018.0

Spalten



umbenannt

	bahnhof_kurz	uic	bahnhof_lang	kanton	bahnhofseigner	jahr
0	AAT	8503124.0	Aathal	ZH	SBB	2018.0
1	ABO	8502000.0	Aarburg-Oftringen	AG	SBB	2018.0
2	AIG	8501400.0	Aigle	VD	SBB	2018.0
3	AIG	8501400.0	Aigle	VD	SBB	2022.0
4	ALT	8506319.0	Altstätten SG	SG	SBB	2018.0
5	ALTD	8503211.0	Altendorf	SZ	SBB	2018.0
6	ALV	8509194.0	Alvaneid	GR	RhR	2018.0

Datenbereinigung – Spalten

Umbenennen & Entfernen



Passagierfrequenz

dtv_tjm_tgm	dwv_tmjo_tfm	dnwv_tmjno_tmgnl	evu_ef_itf
740.0	800.0	610.0	SBB
2500.0	3000.0	1300.0	SBB
6800.0	7700.0	5000.0	RegionAlps, SBB
9100.0	9700.0	7700.0	SBB
2300.0	2700.0	1500.0	SBB, SOB, Turbo
850.0	980.0	560.0	SBB
50.0	50.0	50.0	RhR

Spalten



umbenannt

durchschnittlicher_täglicher_verkehr	durchschnittlicher_werktäglich_verkehr	durchschnittlicher_nicht_werktäglich_verkehr	einbezogene_bahnunternehmen
740.0	800.0	610.0	SBB
2500.0	3000.0	1300.0	SBB
6800.0	7700.0	5000.0	RegionAlps, SBB
9100.0	9700.0	7700.0	SBB
2300.0	2700.0	1500.0	SBB, SOB, Turbo
850.0	980.0	560.0	SBB
50.0	50.0	50.0	RhR

Datenbereinigung – Spalten

Umbenennen & Entfernen



Spalten



Passagierfrequenz

bemerkungen	remarques	note	remarks	geopos	lod
NaN	NaN	NaN	NaN	47.33595913383636, 8.765625135347076	http://lod.opentransportdata.swiss/didok/8503124
NaN	NaN	NaN	NaN	47.320268469494984, 7.908222606719325	http://lod.opentransportdata.swiss/didok/8502000
Ohne TPC.	Sans TPC.	Senza TPC.	Without TPC.	46.31685541859803, 6.9636832118614045	http://lod.opentransportdata.swiss/didok/8501400
Ohne TPC.	Sans TPC.	Senza TPC.	Without TPC.	46.31685541859803, 6.9636832118614045	http://lod.opentransportdata.swiss/didok/8501400
NaN	NaN	NaN	NaN	47.374234807062585, 9.556519883564427	http://lod.opentransportdata.swiss/didok/8506319
NaN	NaN	NaN	NaN	47.19396706593199, 8.822905848209773	http://lod.opentransportdata.swiss/didok/8503211

Spalten



entfernt

bemerkungen	geopos	link
NaN	47.33595913383636, 8.765625135347076	http://lod.opentransportdata.swiss/didok/8503124
NaN	47.320268469494984, 7.908222606719325	http://lod.opentransportdata.swiss/didok/8502000
Ohne TPC.	46.31685541859803, 6.9636832118614045	http://lod.opentransportdata.swiss/didok/8501400
Ohne TPC.	46.31685541859803, 6.9636832118614045	http://lod.opentransportdata.swiss/didok/8501400
NaN	47.374234807062585, 9.556519883564427	http://lod.opentransportdata.swiss/didok/8506319
NaN	47.19396706593199, 8.822905848209773	http://lod.opentransportdata.swiss/didok/8503211

Datenbereinigung – Spalten Zerlegen



Spalten



Passagierfrequenz

bemerkungen	geopos	link
NaN	47.33595913383636, 8.765625135347076	http://lod.opentransportdata.swiss/didok/8503124
NaN	47.320268469494984, 7.908222606719325	http://lod.opentransportdata.swiss/didok/8502000
Ohne TPC.	46.31685541859803, 6.9636832118614045	http://lod.opentransportdata.swiss/didok/8501400
Ohne TPC.	46.31685541859803, 6.9636832118614045	http://lod.opentransportdata.swiss/didok/8501400
NaN	47.374234807062585, 9.556519883564427	http://lod.opentransportdata.swiss/didok/8506319
NaN	47.19396706593199, 8.822905848209773	http://lod.opentransportdata.swiss/didok/8503211

Spalten

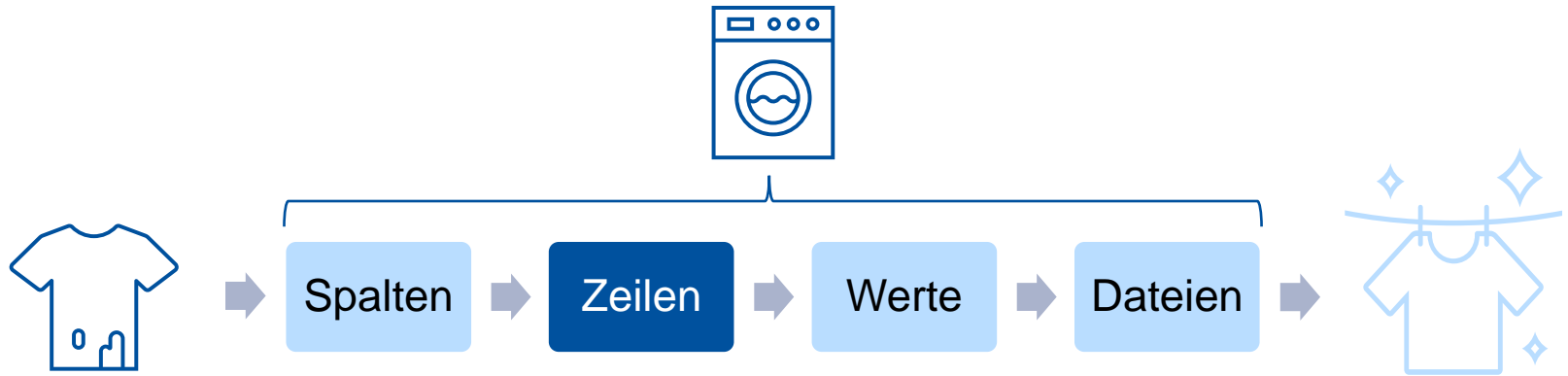


zerteilt

bemerkungen	latitude	longitude	link
NaN	47.33595913383636	8.765625135347076	http://lod.opentransportdata.swiss/didok/8503124
NaN	47.320268469494984	7.908222606719325	http://lod.opentransportdata.swiss/didok/8502000
Ohne TPC.	46.31685541859803	6.9636832118614045	http://lod.opentransportdata.swiss/didok/8501400
Ohne TPC.	46.31685541859803	6.9636832118614045	http://lod.opentransportdata.swiss/didok/8501400
NaN	47.374234807062585	9.556519883564427	http://lod.opentransportdata.swiss/didok/8506319
NaN	47.19396706593199	8.822905848209773	http://lod.opentransportdata.swiss/didok/8503211

Datenbereinigung

Übersicht



Datenbereinigung – Zeilen

Duplikate entfernen



Zeilen



Gleiche Zeilen existieren nur in Fahrzeiten: 12130 Zeilen wurden 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

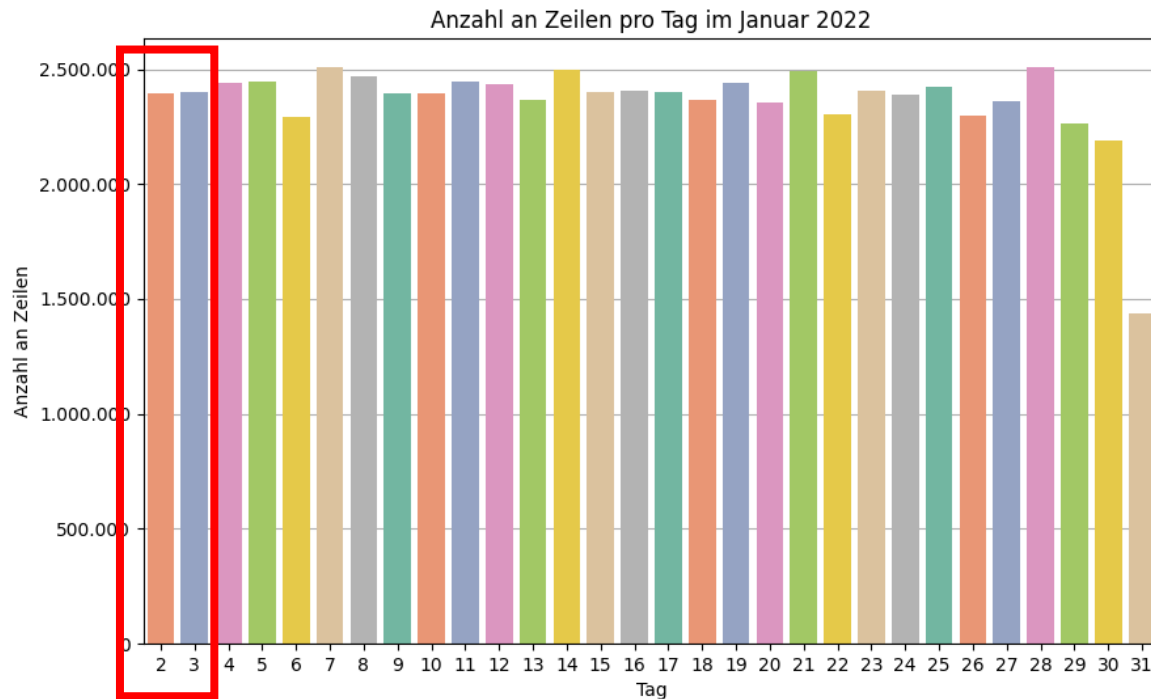
	linie	richtung	betriebsdatum	fahrzeug	kurs	seq_von	halt_diva_von	halt_punkt_diva_von	halt_kurz_von1	datum_von	soll_an_von
0	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	22140
1	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	25680
2	314	1	13.01.22	11443	2	1	657.0	50.0	BDIE	13.01.22	29280
3	314	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 – Zeilen

Fahrzeiten in richtige CSV packen



Zeilen

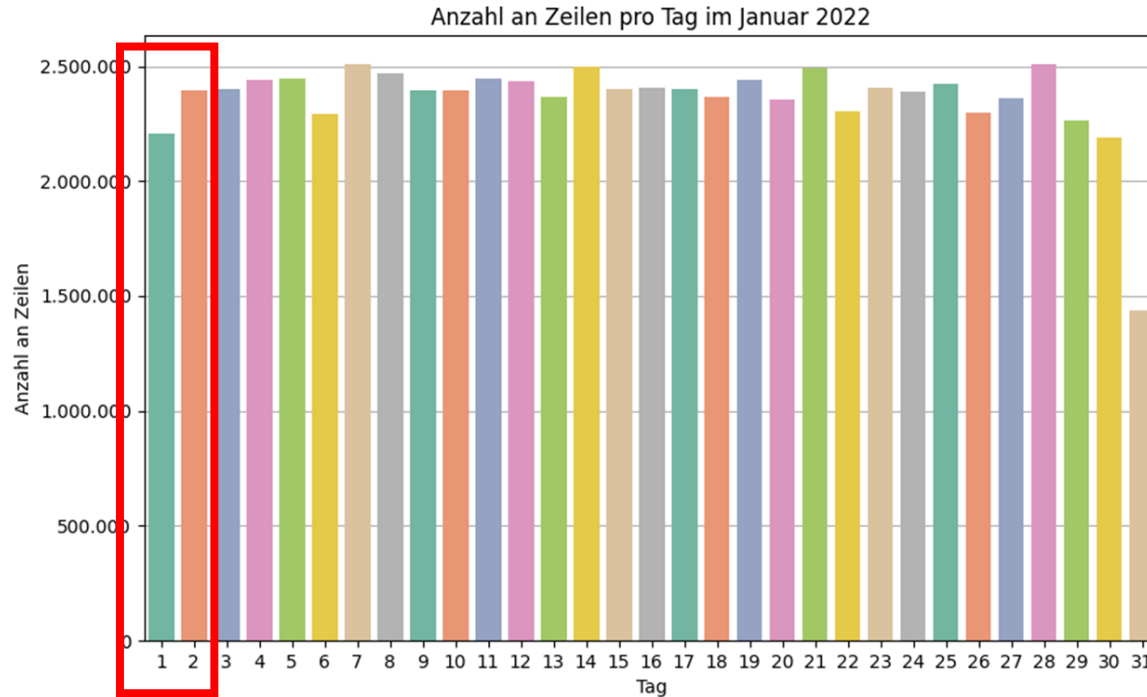


Datenbereinigung – Zeilen

Fahrzeiten in richtige CSV packen

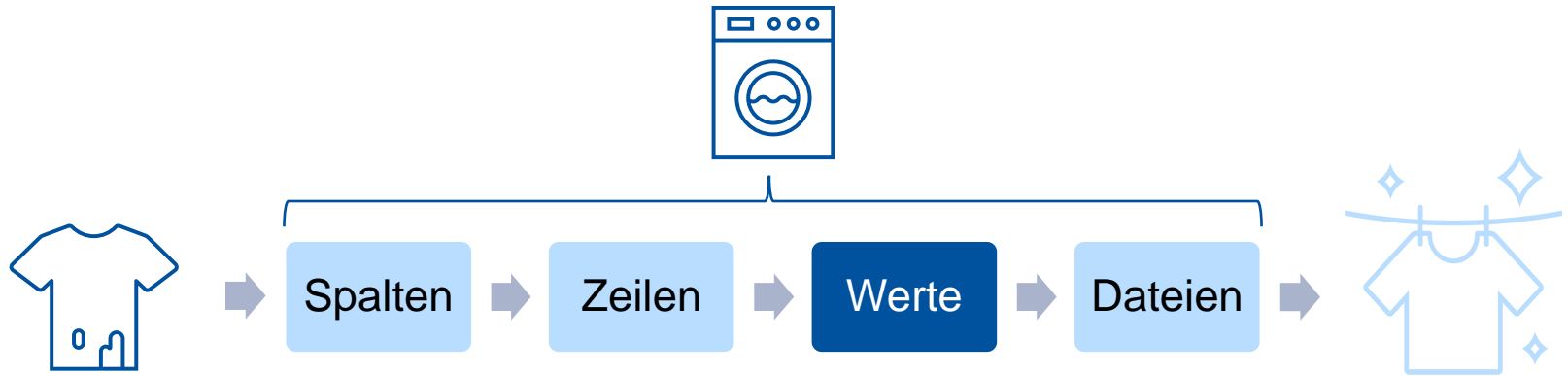


Zeilen



Datenbereinigung

Übersicht



Datenbereinigung – Werte

Leere Werte behandeln



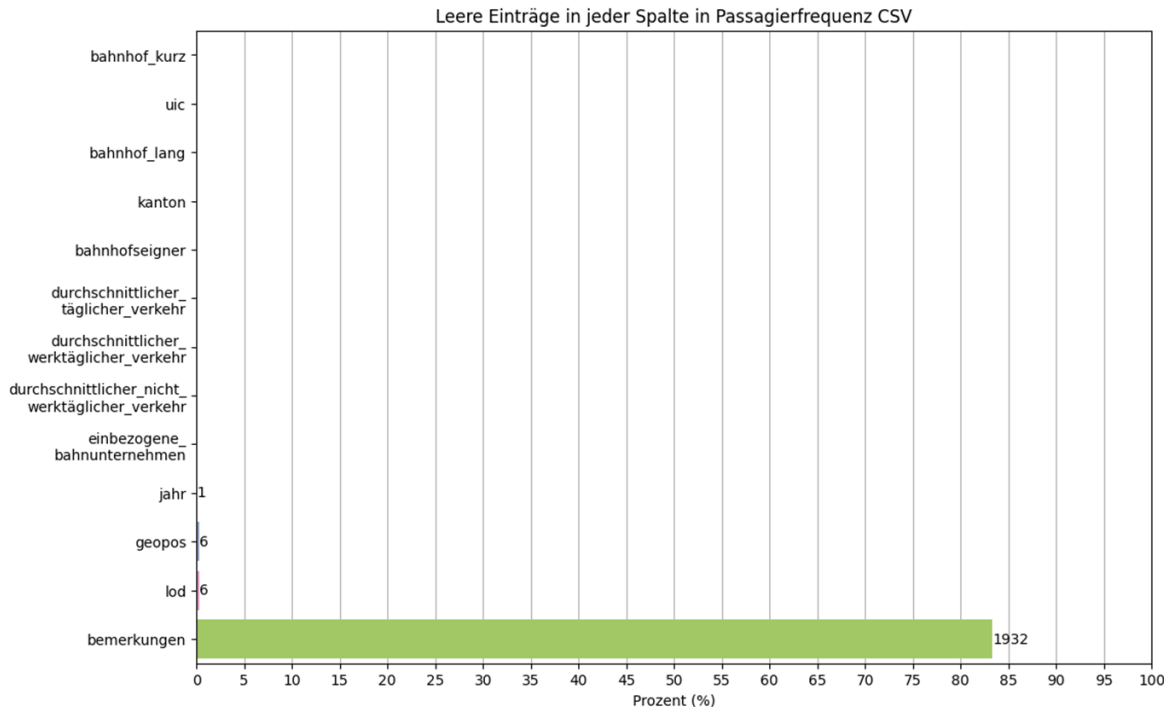
Werte



Passagierfrequenz

Was fehlt?

- 1-mal Jahr
 - Manuell hinzugefügt
 - Immer 2 Jahre pro Haltestelle
 - Klar, welcher Wert rein kommt
- Je 6-mal Latitude, Longitude, Link
 - Manuell hinzugefügt
 - Mit uic Haltestelleninformationen gefunden
- >80 % Bemerkungen
 - Muss nichts getan werden



Datenbereinigung – Werte

Leere Werte behandeln



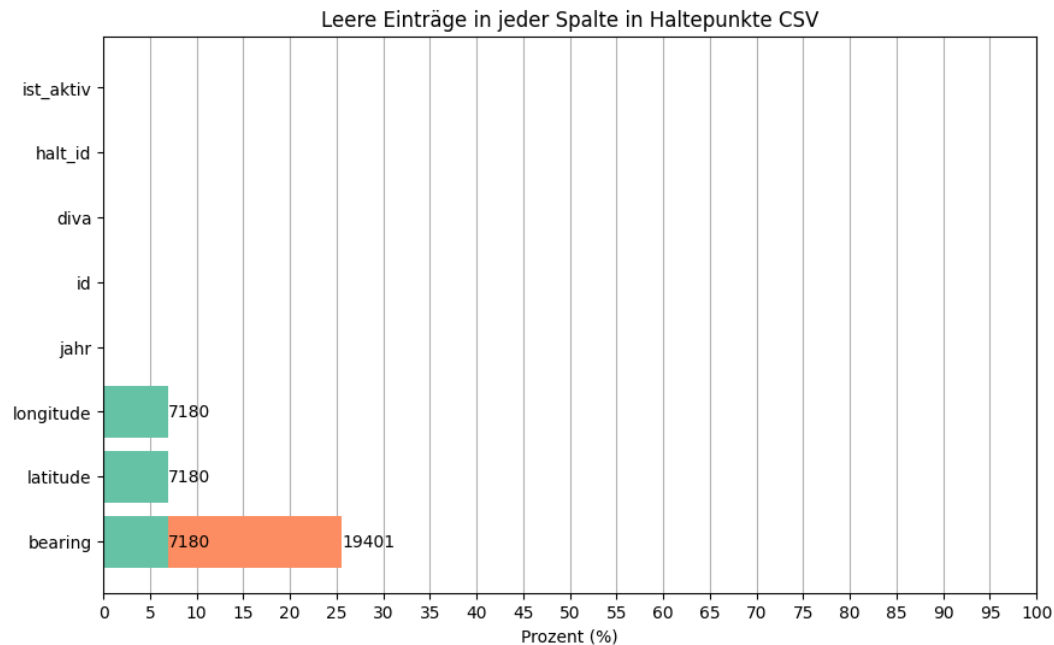
Werte



Haltepunkte

Was fehlt?

- Bearing fehlt, aber Latitude und Longitude gibt es (braun)
 - Bearing = -1 gesetzt
- Latitude, Longitude und Bearing fehlen gleichzeitig (gelb)
 - Keine akkuraten GPS-Daten gefunden
 - Aber...



Datenbereinigung – Werte

Leere Werte behandeln



Haltepunkte & Haltestellen

- Haltestellen GPS Daten gefunden (Excel Datei & Google Maps)
 - Diese in Haltestellen hinzugefügt
 - Automatisch und teilweise manuell
 - Jetzt haben >99% Haltestellen GPS Daten
 - Rest hat -1 als Latitude und Longitude bekommen (<1%)
- Ein Haltepunkt gehört zu einer Haltestelle
 - Haltepunkte ohne GPS-Daten haben, die von Haltestellen bekommen
 - Bearing = -2, um zu wissen, dass GPS-Daten von Haltepunkt nicht 100% akkurat sind
 - Jetzt haben 100% Haltepunkte GPS-Daten

Datenbereinigung – Werte

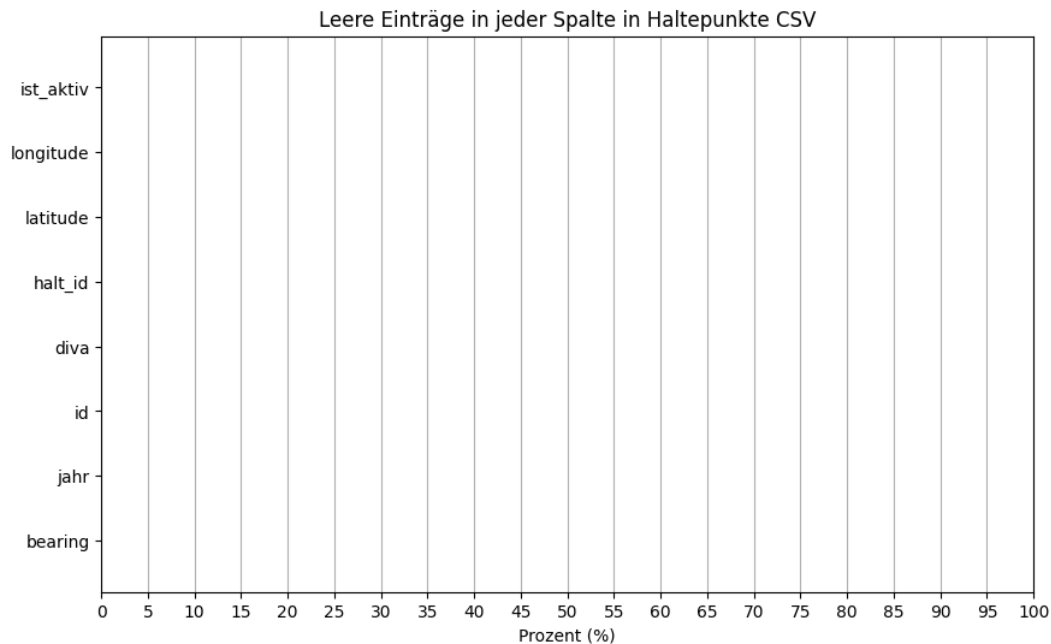
Leere Werte behandeln



Werte



Haltepunkte



Datenbereinigung – Werte

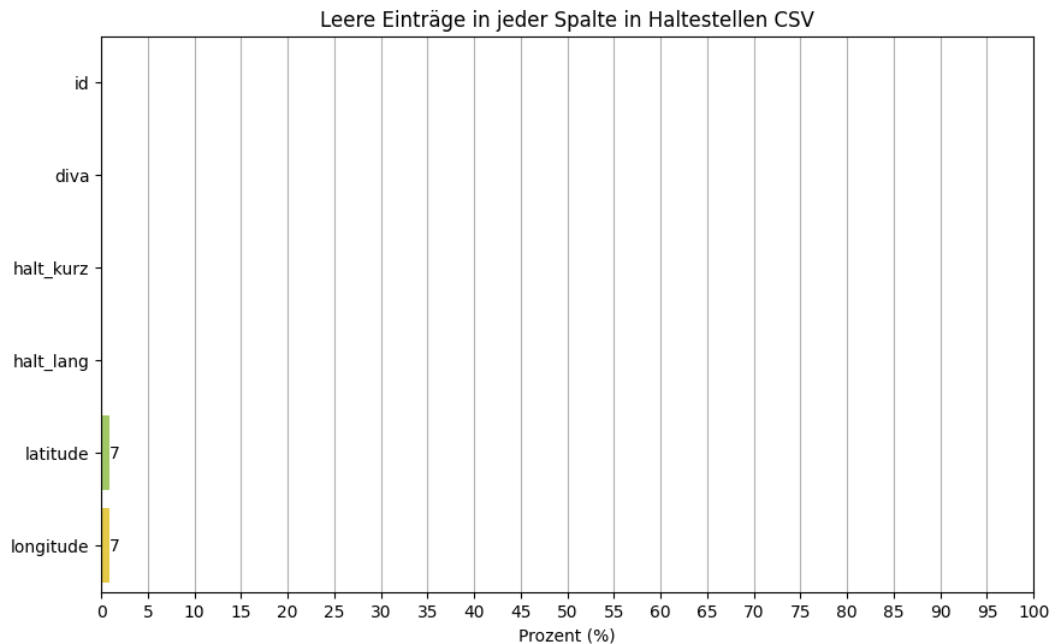
Leere Werte behandeln



Werte

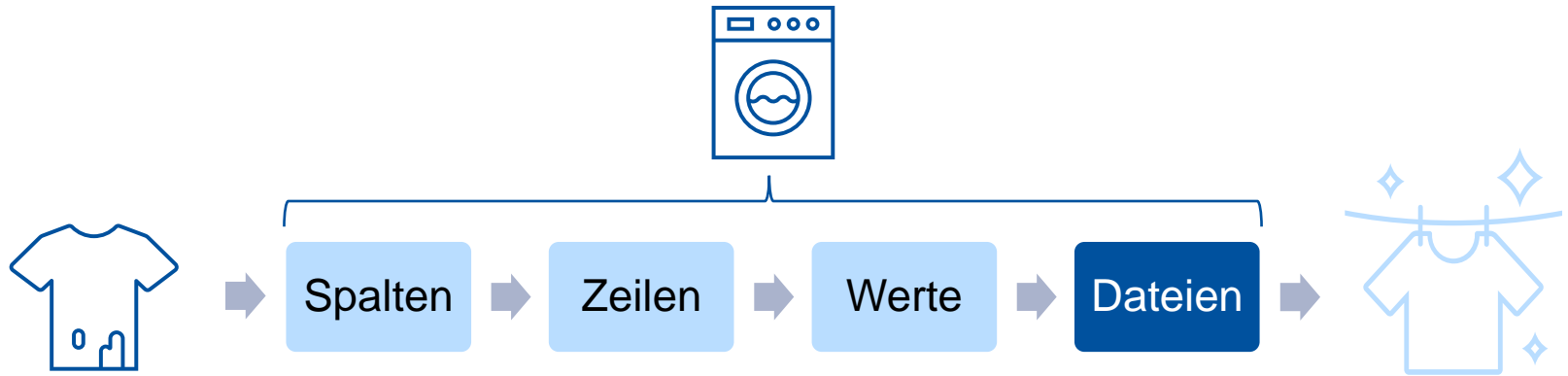


Haltestellen



Datenbereinigung

Übersicht



Datenbereinigung – Touch Ups

Csv Dateien zusammenfügen



Haltestellen

- Haben alle Haltestellen dieselben Einträge in jedem Jahr?
 - Fast! Haltestellen Namen (halt_lang und halt_kurz) wurden manchmal geändert
- Sind alle Haltestellen von allen Jahren auch im letzten Jahr?
 - Ja
- Haltestellen csv's wurden gelöscht, nur die 2022 csv wurde behalten

Datenbereinigung – Touch Ups

Csv Dateien zusammenfügen

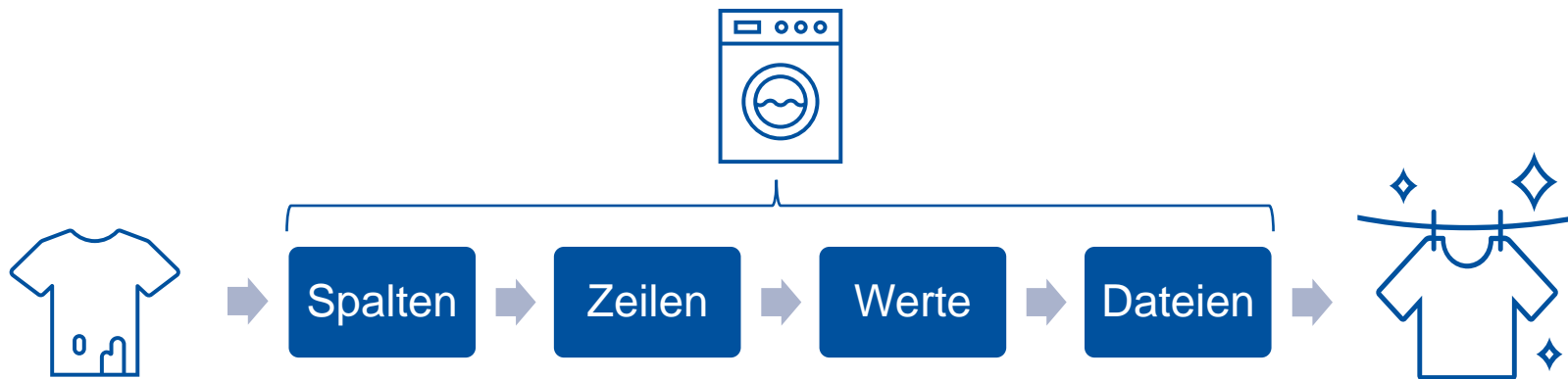


Haltepunkte

- Haltepunkte csv's wurden alle in eine csv Datei zusammen gefügt
 - Jahr Spalte wurde hinzugefügt

Datenbereinigung

Übersicht



Zum Besprechen

- Es gibt Haltepunkte, mit denselben Einträgen
 - Jahr, Fremdschlüssel zu Haltestellen, GPS-Koordinate, ist_aktiv, etc. gleich
 - Aber verschiedene Haltepunkt id
 - Vorschlag für Bereinigung:
 - Doppelte Einträge **löschen**, nur eins behalten:
 - Verweise von Fahrzeiten zu diesen Doppelten Haltepunkten nur zu einem von den Haltepunkten ändern
 - Doppelte Haltepunkte löschen
 - Oder...
 - So **belassen**

Zum Besprechen

- Es gibt Fahrzeiten, mit Fremdschlüsseln zu Haltepunkten, die nicht existieren
 - 4865 Zeilen, 6 verschiedene Haltepunkt id's
 - Vorschlag für Bereinigung:
 - Haltepunkte **ersetzen** von der Haltestelle:
 - Fahrzeiten haben auch Haltestellen id's (aber kein offizieller Fremdschlüssel)
 - Haltestellen haben Haltepunkte zugewiesen
 - Einen von diesen Haltepunkten mit den nicht existierenden Haltepunkt id's ersetzen
 - Oder...
 - Haltepunkte **hinzufügen** von der Haltestelle:
 - Diese fehlenden Haltepunkte in Haltepunkten Tabelle einfügen
 - GPS-Daten von Haltestelle nehmen, auf die die Fahrzeiten verweisen (wieder kein offizieller Fremdschlüssel)

Zeitaufwand



- Ozan: ~83 Stunden
 - Bereinigung
 - Programmierung
 - Manuelle Bereinigung
 - Visualisierungen
 - Präsentationsfolien
- Tony: ~20 Stunden
 - Manuelle Bereinigung
 - Visualisierungen
- Insgesamt: ~103 Stunden