EDA Dataset & Model

	date_event	type	value	city_id
6	2023-07-04 22:12:37	HEAVY	0.020354	1
7	2023-07-04 22:12:39	HEAVY	0.105000	1
8	2023-07-07 09:14:57	HEAVY	0.151063	1
11	2023-07-10 06:38:05	HEAVY	0.151063	1
12	2023-07-10 12:25:05	HEAVY	0.020354	1
...	...	...	...	...
58747	2023-07-24 05:39:27	LIGHT	0.308120	1
58750	2023-07-24 05:40:02	MODERATE	0.238860	1
58752	2023-07-25 15:59:25	MODERATE	0.301226	1
58753	2023-07-25 15:59:26	MODERATE	0.378487	1
58755	2023-07-25 16:09:49	MODERATE	0.170212	1

21881 rows × 4 columns

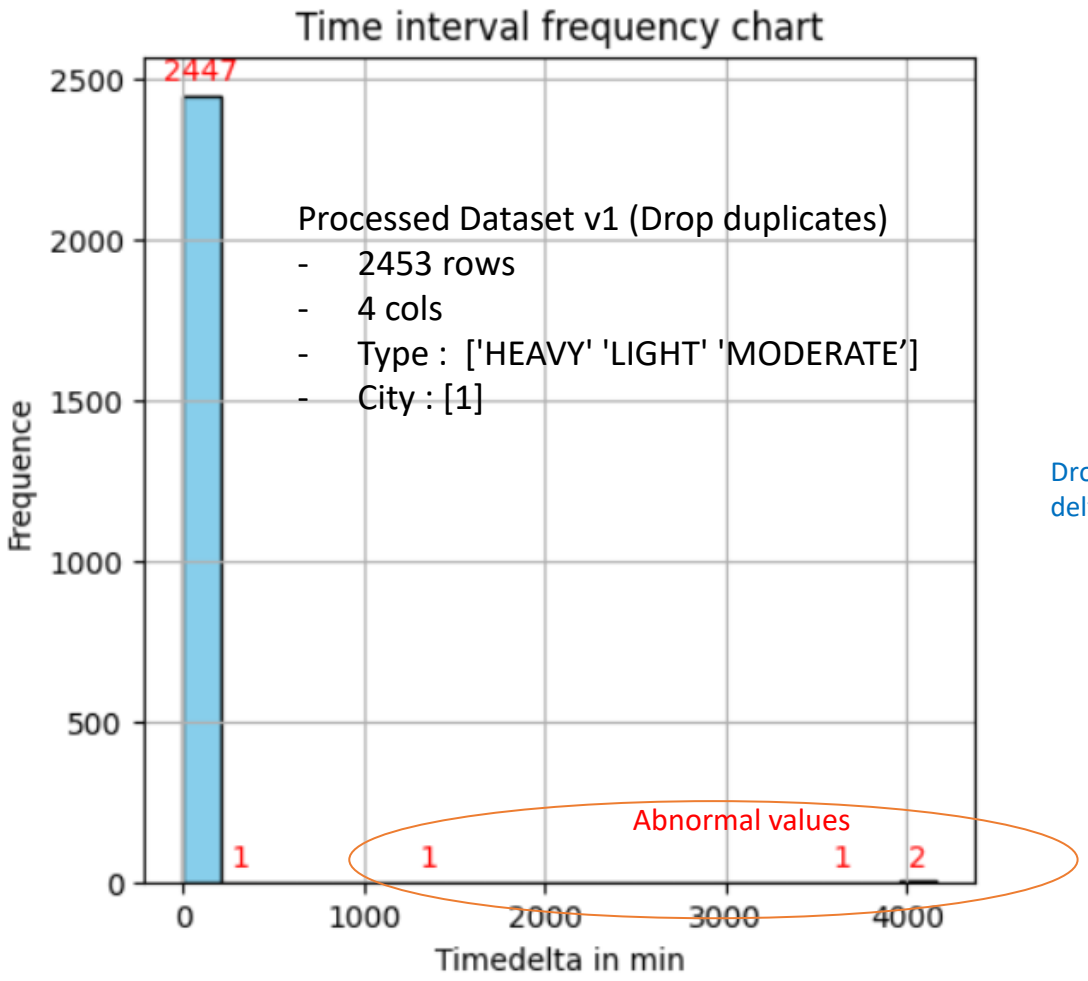
Raw Dataset
- 21881 rows
- 4 cols
- Type : ['HEAVY' 'LIGHT' 'MODERATE']
- City : [1]

	date_event	type	value	city_id
6	2023-07-04 22:12:37	HEAVY	0.020354	1
7	2023-07-04 22:12:39	HEAVY	0.105000	1
8	2023-07-07 09:14:57	HEAVY	0.151063	1
11	2023-07-10 06:38:05	HEAVY	0.151063	1
12	2023-07-10 12:25:05	HEAVY	0.020354	1
...	...	...	...	...
56138	2023-07-26 06:33:53	HEAVY	0.253121	1
56146	2023-07-26 06:44:22	MODERATE	0.274233	1
39441	2023-07-26 07:02:49	HEAVY	0.378487	1
39453	2023-07-26 07:02:53	MODERATE	0.528919	1
39445	2023-07-26 07:13:19	LIGHT	0.274233	1

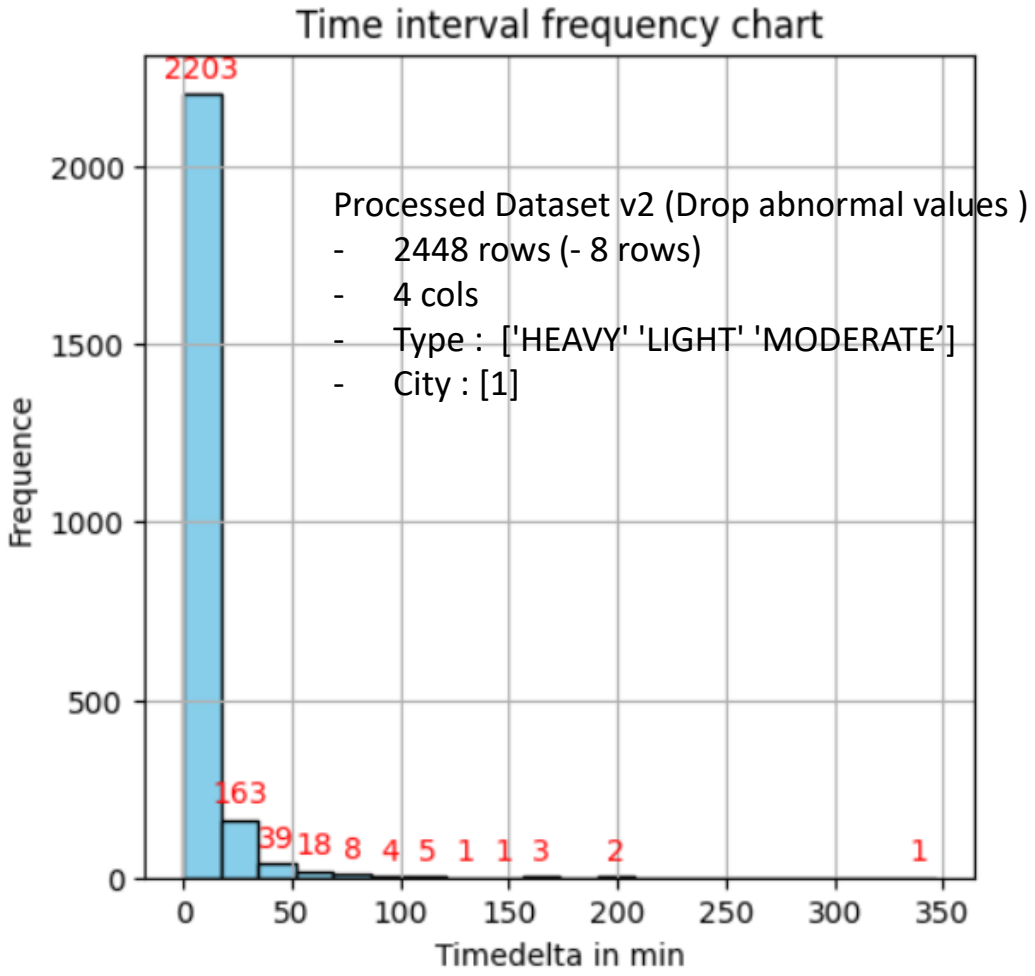
2453 rows × 4 columns

Processed Dataset v1 (drop duplicate)
- 2453 rows (- 88.8%)
- 4 cols
- Type : ['HEAVY' 'LIGHT' 'MODERATE']
- City : [1]
- Date : 7/4 – 7/26

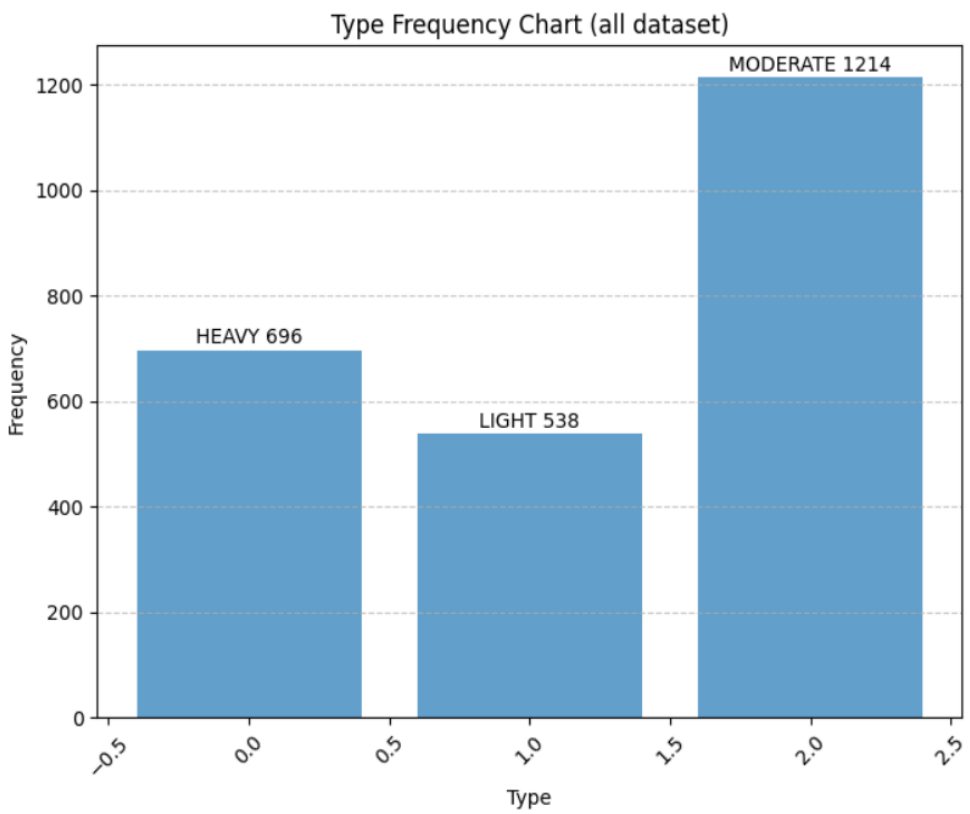
Time interval frequency chart (in min)

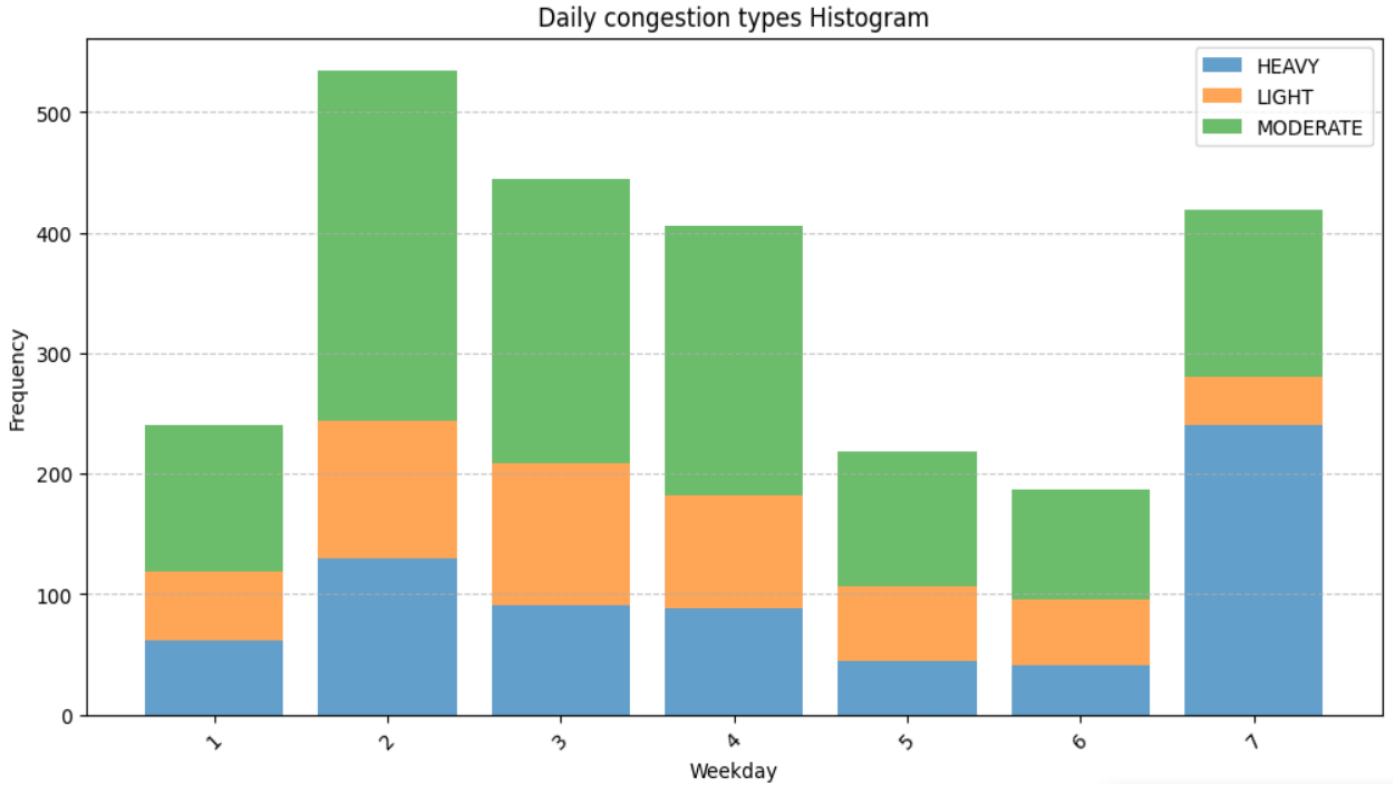


Drop abnormal values
delta>12*60 min

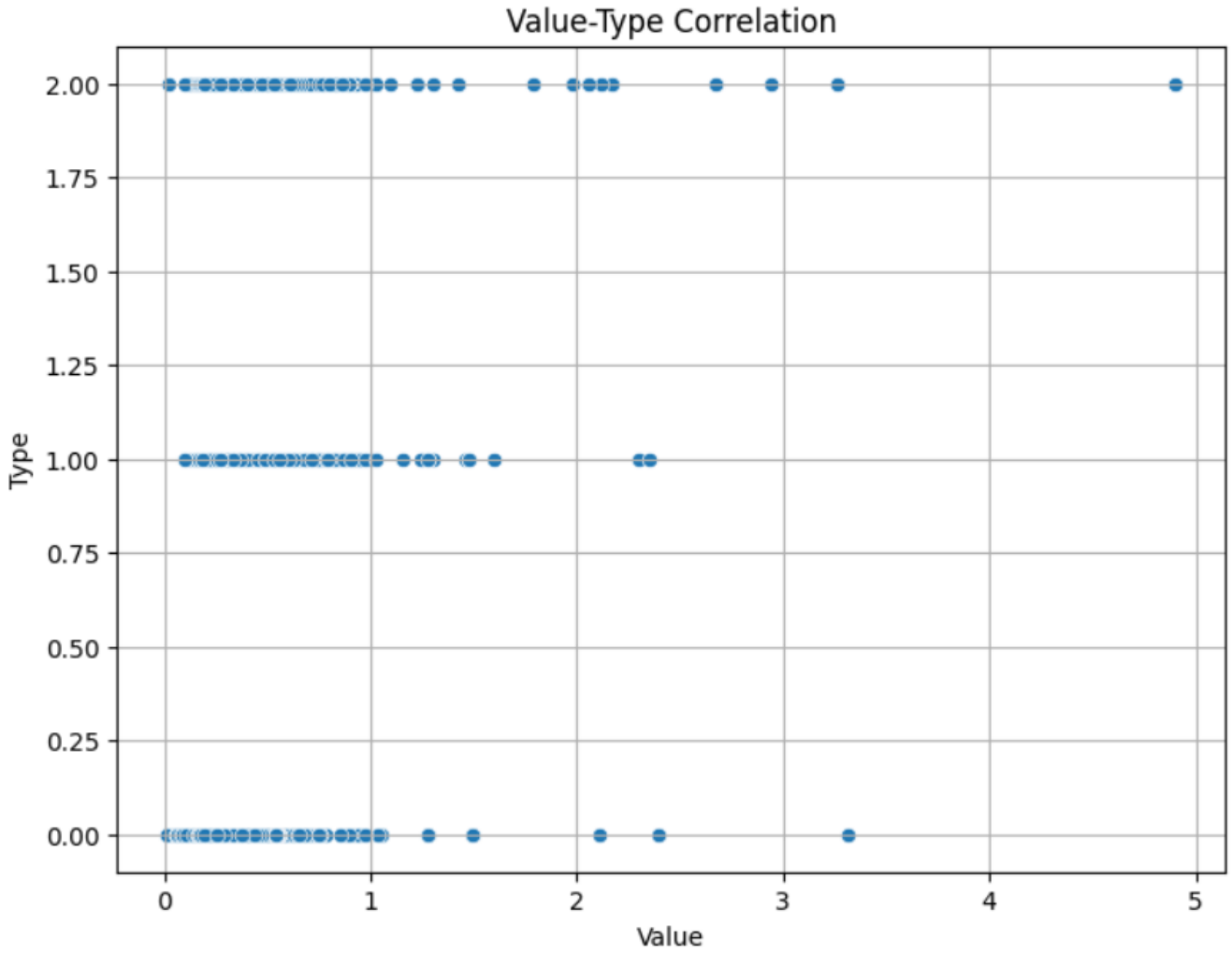


Type Frequency Chart

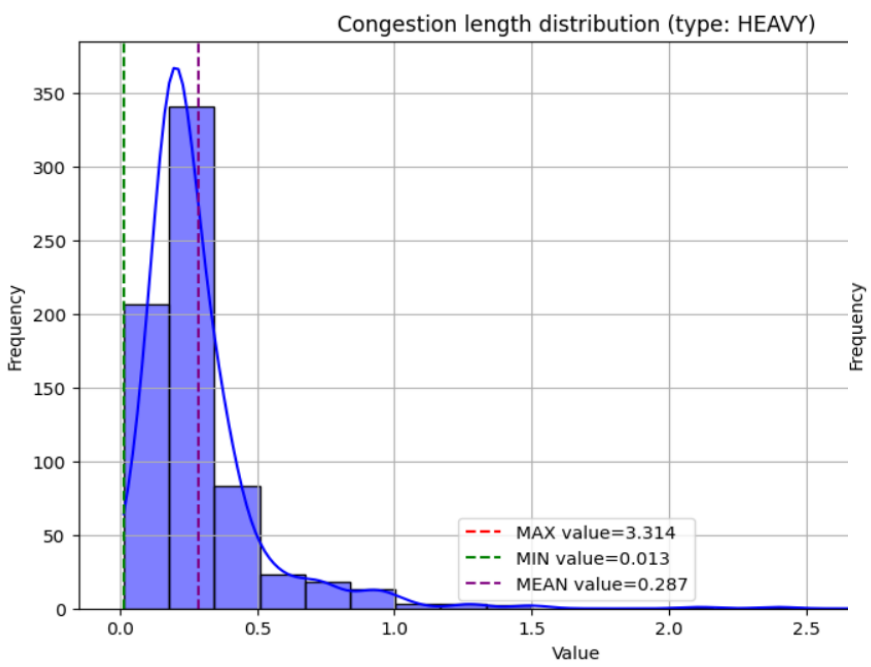


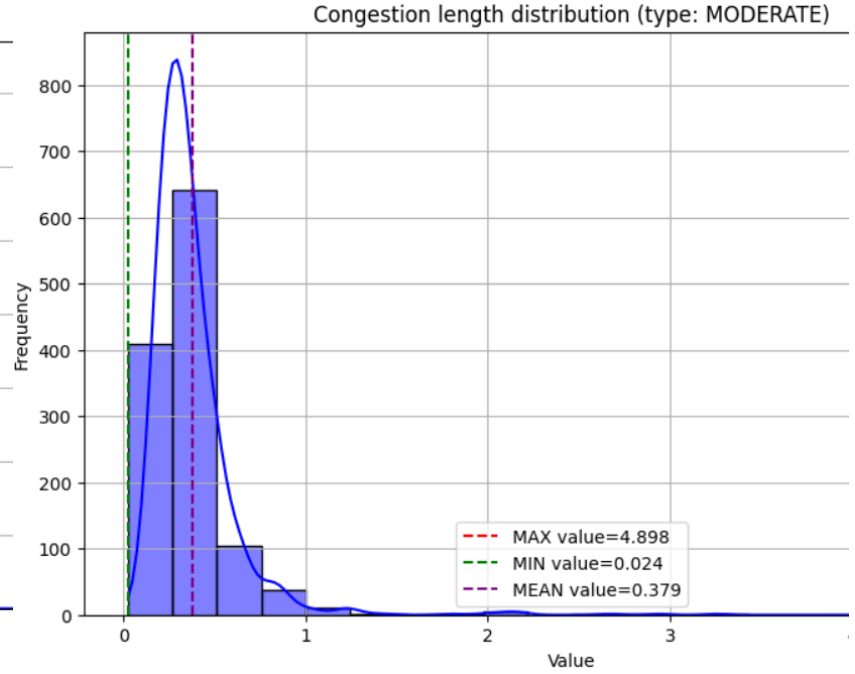


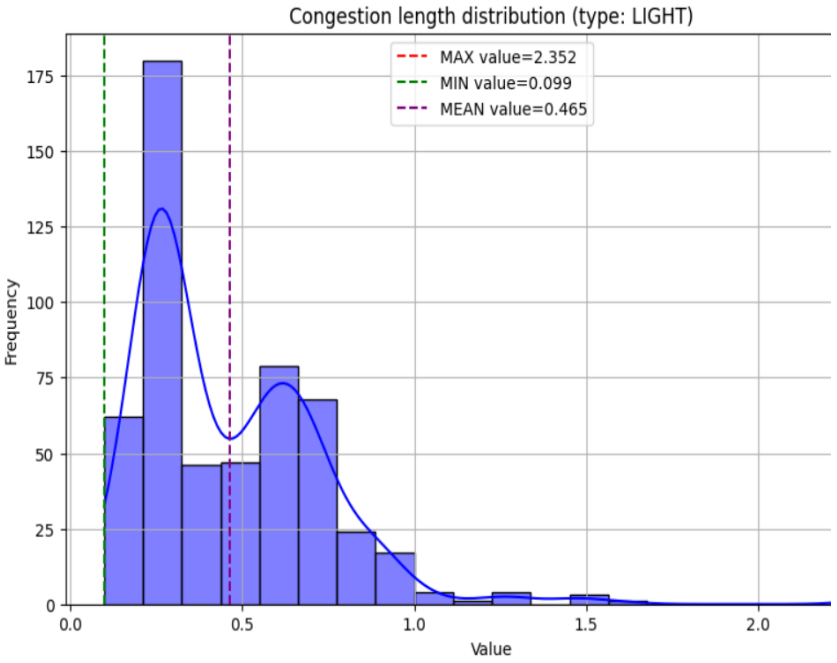
Value-type correlation



Congestion length distribution chart







Dataset from 7/11 for city1

	hour_index	weekday	value
0	0	2.0	0.187787
1	1	2.0	0.404619
2	2	2.0	0.441498
3	3	2.0	0.724283
4	4	2.0	0.193599
...	...	...	...
266	347	3.0	0.218163
267	348	3.0	0.558570
268	352	3.0	0.262792
269	353	3.0	0.315643
270	354	3.0	0.393880

Time intervals are different → need a unified Time interval

- hour index
- Cumulated hour (ref : the first timestmp)
- Calculate mean-value in one hour index
- We have missing values → 270 rows but 354 hours

	hour_index	mean_value	day
0	0	0.187787	2
1	1	0.404619	2
2	2	0.441498	2
3	3	0.724283	2
4	4	0.193599	2
...	...	...	...
349	349	0.000000	3
350	350	0.000000	3
351	351	0.000000	3
352	352	0.262792	3
353	353	0.315643	3

354 rows × 3 columns

Padding missing values by 0

- We have 354 rows and 354 hours

Dataset

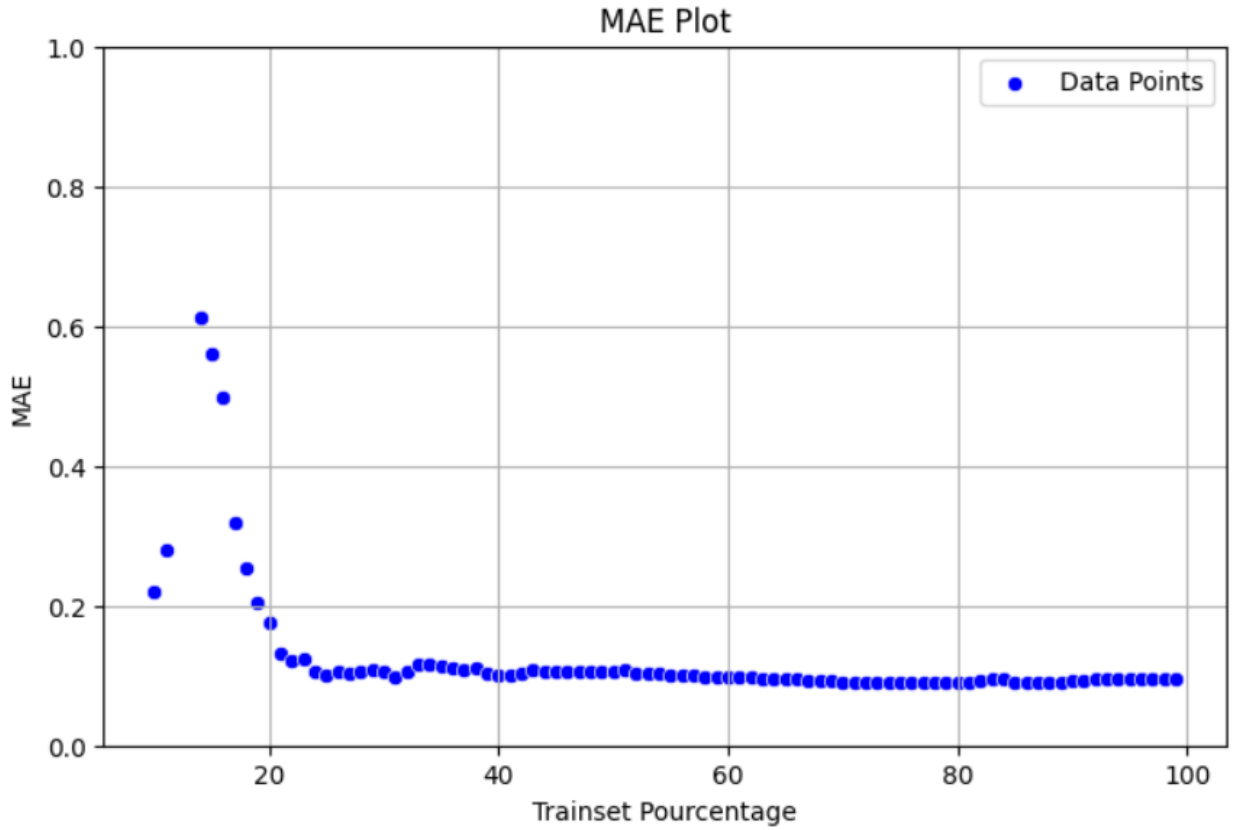
dex	mean_value	value_shift_1	value_shift_2	value_shift_3	value_shift_4	value_shift_5	value_shift_6	value_shift_7	...	value_shift_22	value_shift_23	value_shift_24
24	0.365277	0.295500	0.333460	0.314829	0.402290	0.389791	0.330464	0.330314	...	0.441498	0.404619	0.187787
25	0.391379	0.365277	0.295500	0.333460	0.314829	0.402290	0.389791	0.330464	...	0.724283	0.441498	0.404619
26	0.440046	0.391379	0.365277	0.295500	0.333460	0.314829	0.402290	0.389791	...	0.193599	0.724283	0.441498
27	0.362597	0.440046	0.391379	0.365277	0.295500	0.333460	0.314829	0.402290	...	0.154464	0.193599	0.724283
28	0.243996	0.362597	0.440046	0.391379	0.365277	0.295500	0.333460	0.314829	...	0.214764	0.154464	0.193599
...	...	...	...	...	...	...	...	...	...	...	...	...
349	0.000000	0.558570	0.218163	0.428380	0.177547	0.212264	0.266966	0.244087	...	0.000000	0.000000	0.396077
350	0.000000	0.000000	0.558570	0.218163	0.428380	0.177547	0.212264	0.266966	...	0.297351	0.000000	0.000000
351	0.000000	0.000000	0.000000	0.558570	0.218163	0.428380	0.177547	0.212264	...	0.523222	0.297351	0.000000
352	0.262792	0.000000	0.000000	0.000000	0.558570	0.218163	0.428380	0.177547	...	0.294434	0.523222	0.297351
353	0.315643	0.262792	0.000000	0.000000	0.000000	0.558570	0.218163	0.428380	...	0.394900	0.294434	0.523222

y

X

AutoRegression
- Use last 24h data (mean value) to predict actual mean value

Linear Regression Model



AutoRegression
- Use last 24h data (mean value) to predict actual mean value

	date_event	type	value	city_id
0	2023-06-05 06:45:02	HEAVY	0.129126	3
1	2023-06-23 12:17:14	HEAVY	0.132907	2
2	2023-06-30 02:46:47	HEAVY	0.126696	3
3	2023-06-05 06:45:02	HEAVY	0.129126	3
4	2023-06-23 12:17:14	HEAVY	0.132907	2
...	...	...	...	...
58751	2023-07-25 15:52:54	MODERATE	0.561016	2
58752	2023-07-25 15:59:25	MODERATE	0.301226	1
58753	2023-07-25 15:59:26	MODERATE	0.378487	1
58754	2023-07-25 16:09:45	MODERATE	0.303189	3
58755	2023-07-25 16:09:49	MODERATE	0.170212	1

58756 rows × 4 columns

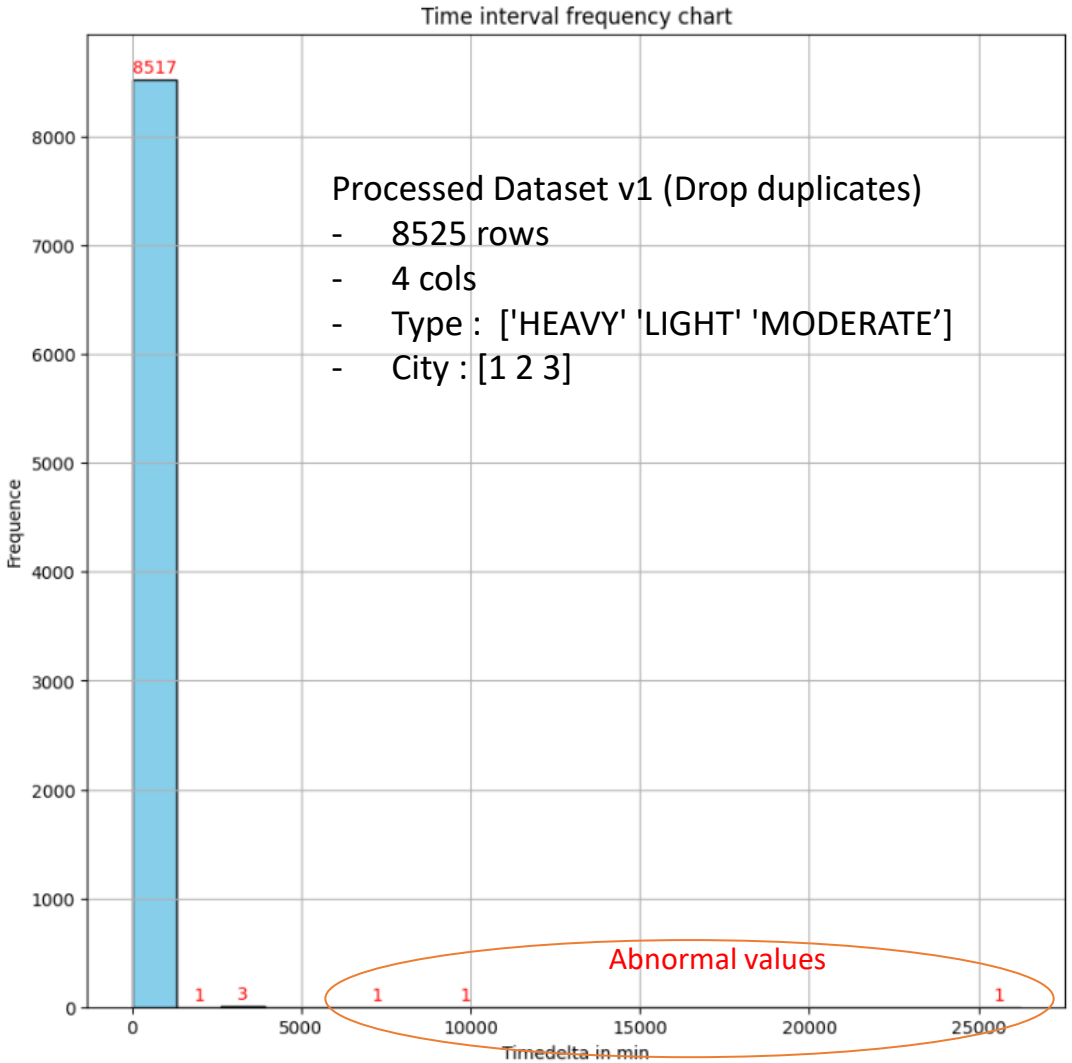
Raw Dataset
- 58756 rows
- 4 cols
- Type : ['HEAVY' 'LIGHT' 'MODERATE’]
- City : [1 2 3]

	date_event	type	value	city_id
0	2023-06-05 06:45:02	HEAVY	0.129126	3
1	2023-06-23 12:17:14	HEAVY	0.132907	2
2	2023-06-30 02:46:47	HEAVY	0.126696	3
6	2023-07-04 22:12:37	HEAVY	0.020354	1
7	2023-07-04 22:12:39	HEAVY	0.105000	1
...	...	...	...	...
39445	2023-07-26 07:13:19	LIGHT	0.274233	1
39446	2023-07-26 07:13:27	LIGHT	0.350618	3
39455	2023-07-26 07:13:29	MODERATE	0.182517	3
39447	2023-07-26 07:13:31	LIGHT	0.203798	3
39448	2023-07-26 07:13:33	LIGHT	0.192763	3

8525 rows × 4 columns

Processed Dataset v1 (drop duplicate)
- 8525 rows (- 85.5%)
- 4 cols
- Type : ['HEAVY' 'LIGHT' 'MODERATE’]
- City : [1 2 3]

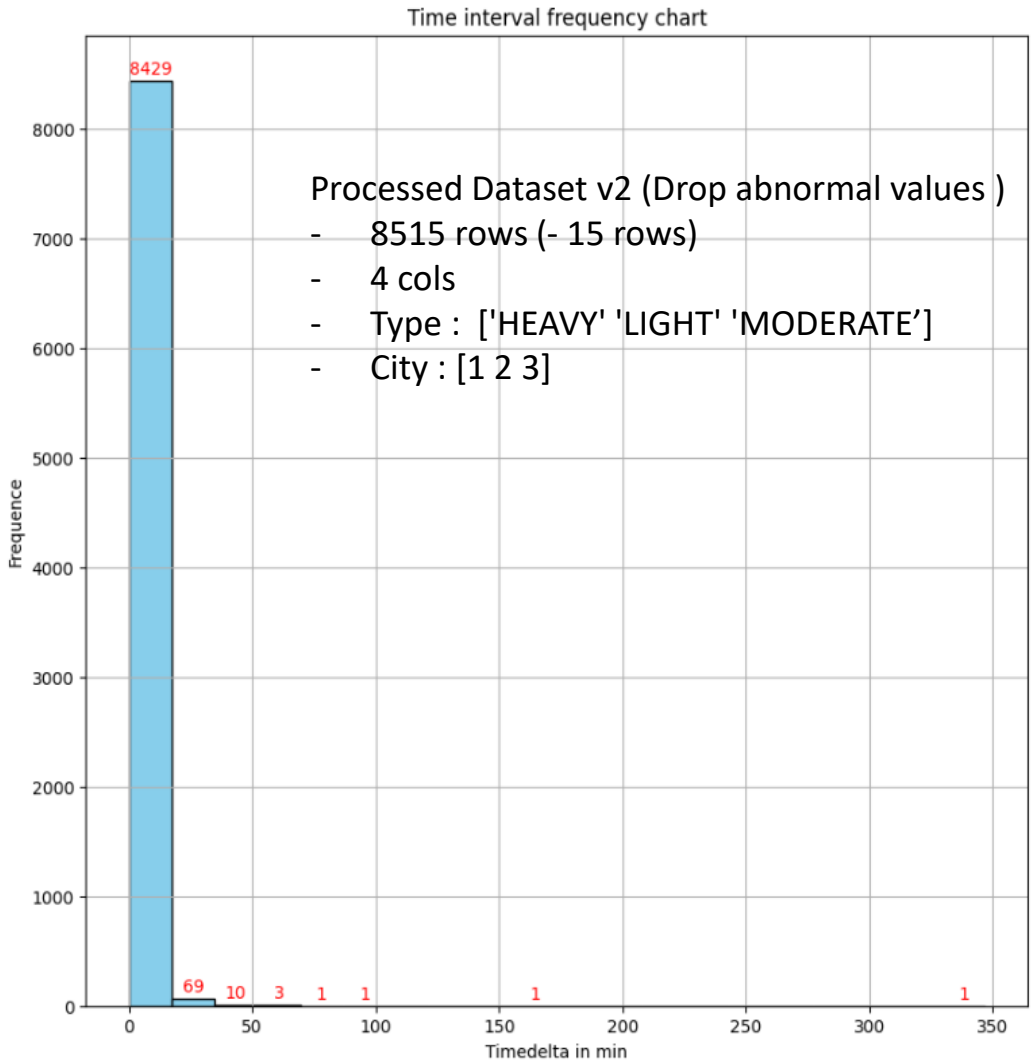
Time interval frequency chart (in min)



Processed Dataset v1 (Drop duplicates)

- 8525 rows
- 4 cols
- Type : ['HEAVY' 'LIGHT' 'MODERATE']
- City : [1 2 3]

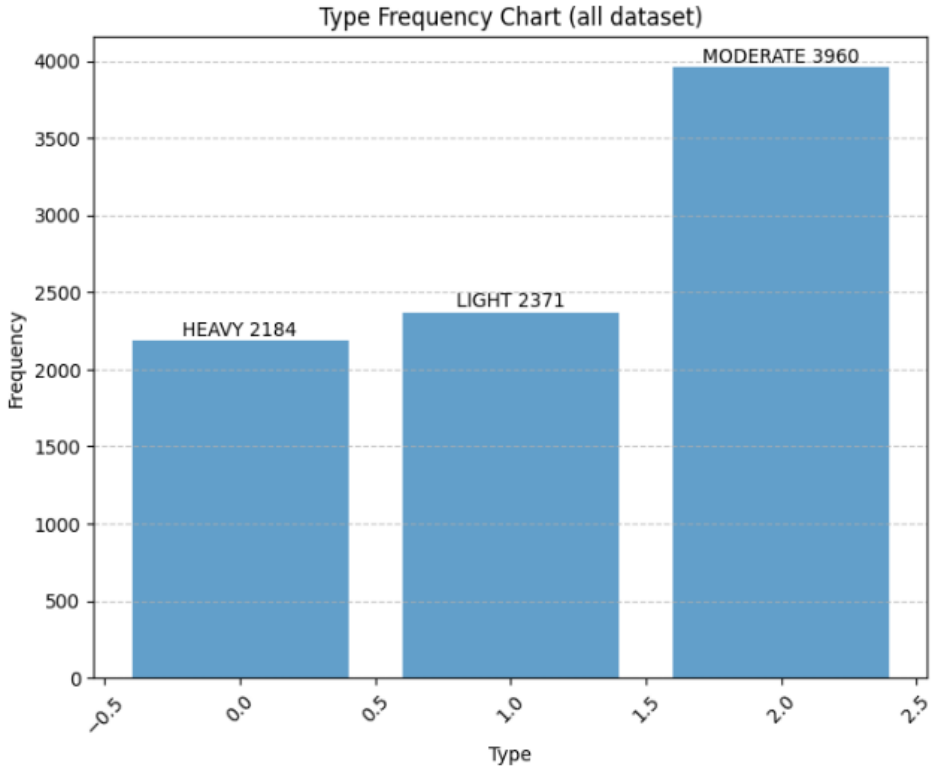
Drop abnormal values
delta>12*60 min

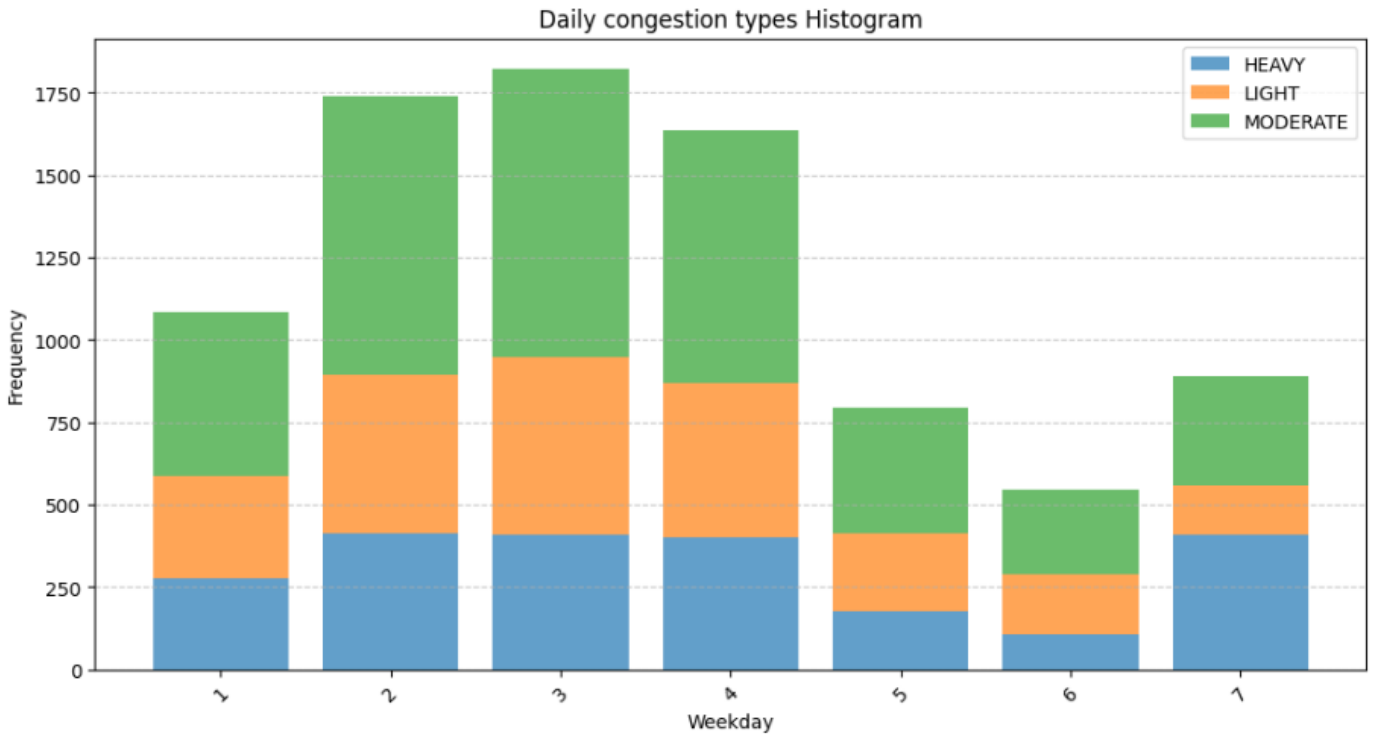


Processed Dataset v2 (Drop abnormal values)

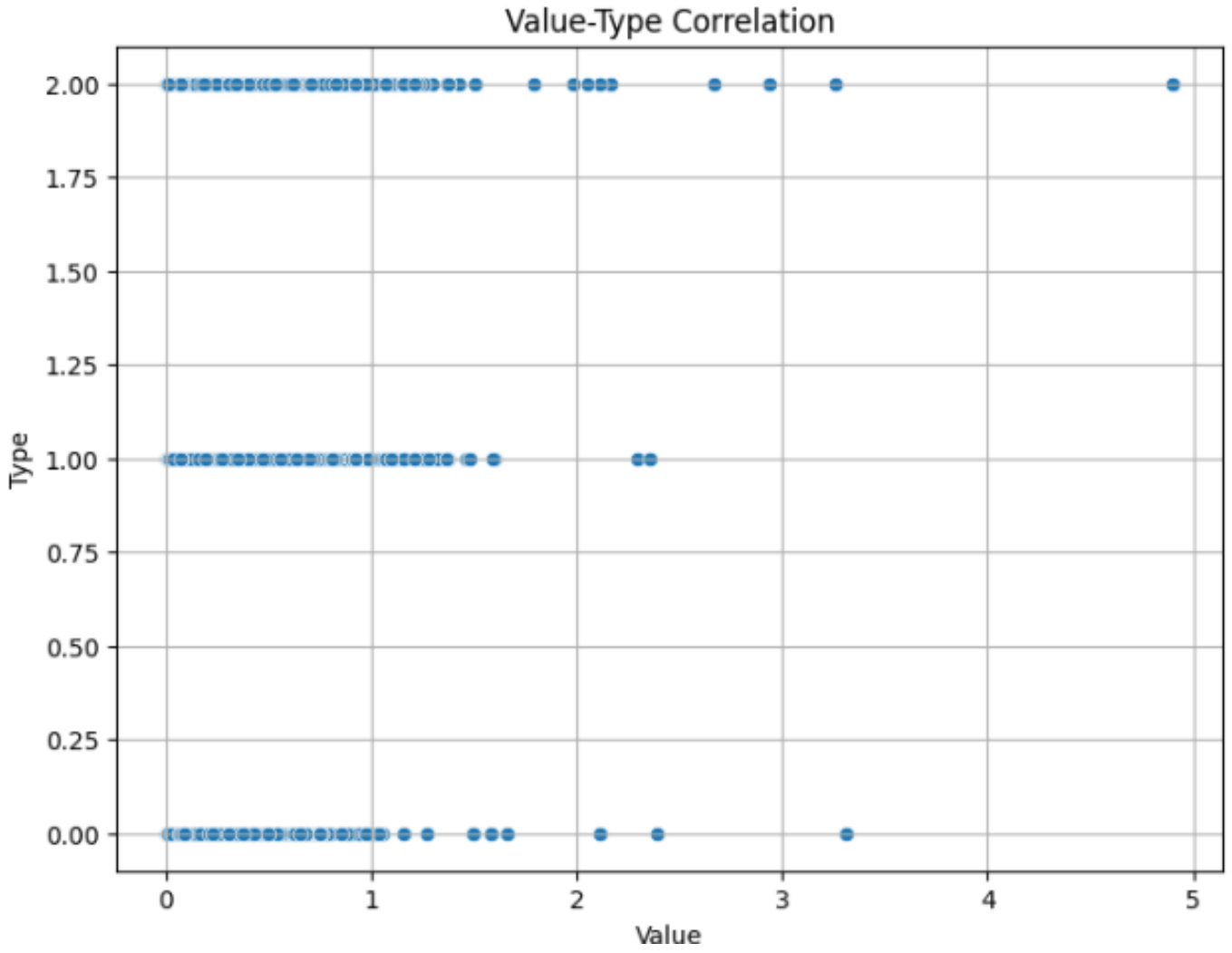
- 8515 rows (- 15 rows)
- 4 cols
- Type : ['HEAVY' 'LIGHT' 'MODERATE']
- City : [1 2 3]

Type Frequency Chart





Value-type correlation



Congestion length distribution chart