

PRACTICAL ASSIGNMENT

WATER CONSUMER CLASSIFICATION

BIP: Machine Learning for Data Science

Group F:

Ayoze Jesús Guanche Núñez
(ULL)

Joran Andringa (DSV)

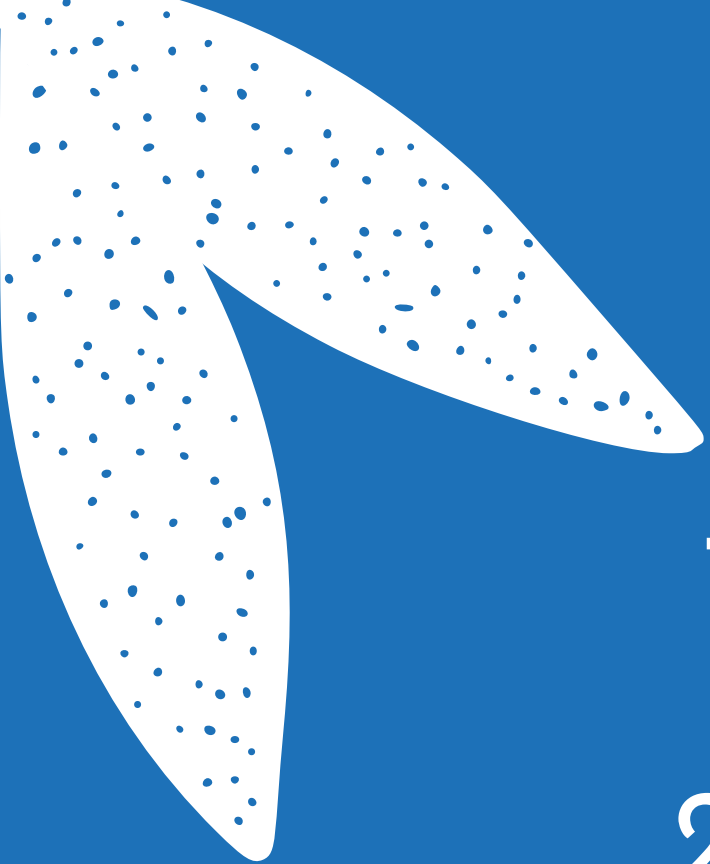
Roberto Chavez Trujillo (ULL)

Barbare Pipia (BTU)

Priscila Reyes Torres (ULL)

Andre Felipe Bianek da Silveira
(VU)





FEATURES

1. YEAR

2. MONTH

3. CONSUMER_TYPE

4. CONSUMPTION

5. CONSUMER_NUMBER

6. INSTALLATION_ZONE





CONSUMER TYPE:

1. DOMESTIC
2. RURAL DOMESTIC
3. INDUSTRIAL
4. RURAL COMMERCIAL
5. CONSTRUCTION
6. LOW INCOME FAMILIES
7. RURAL EXPANSION

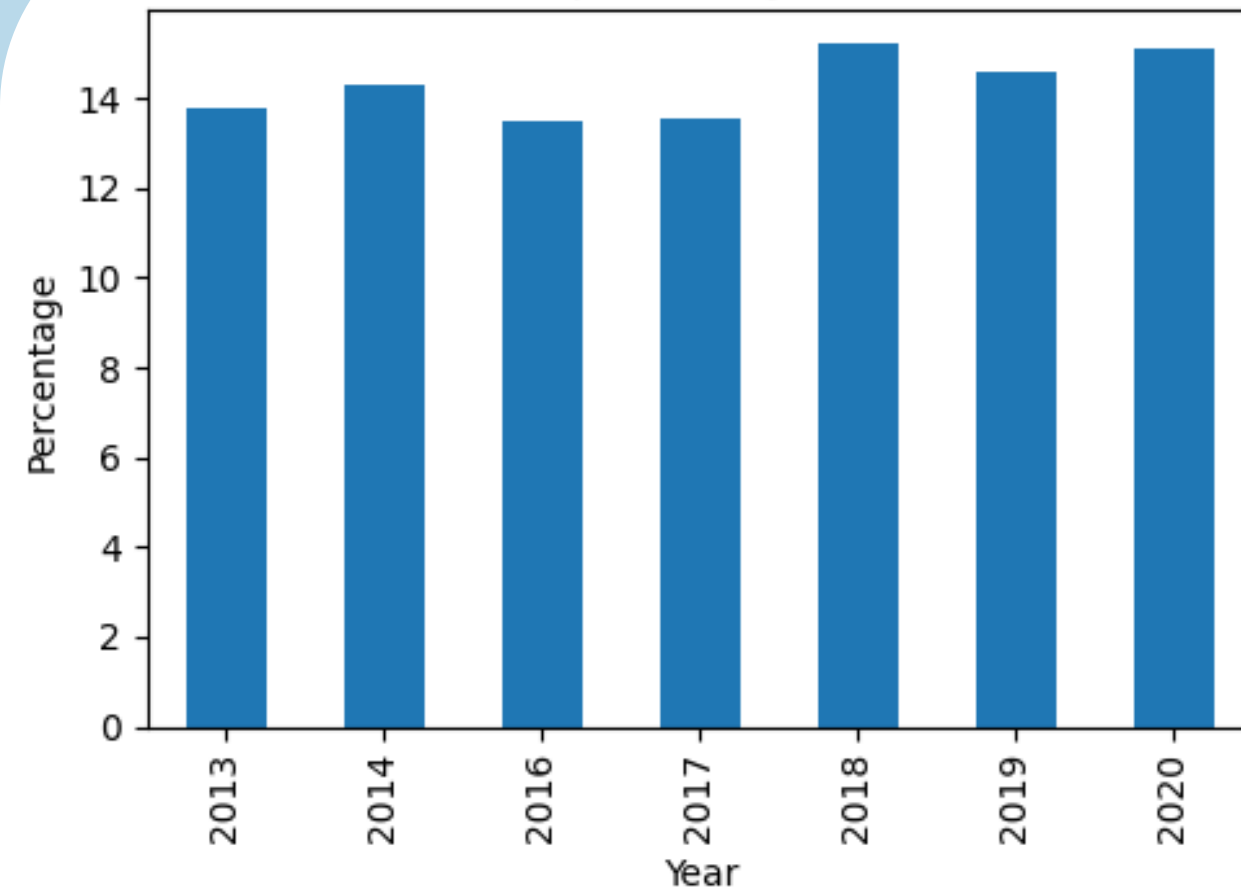


DATASET DESCRIPTION

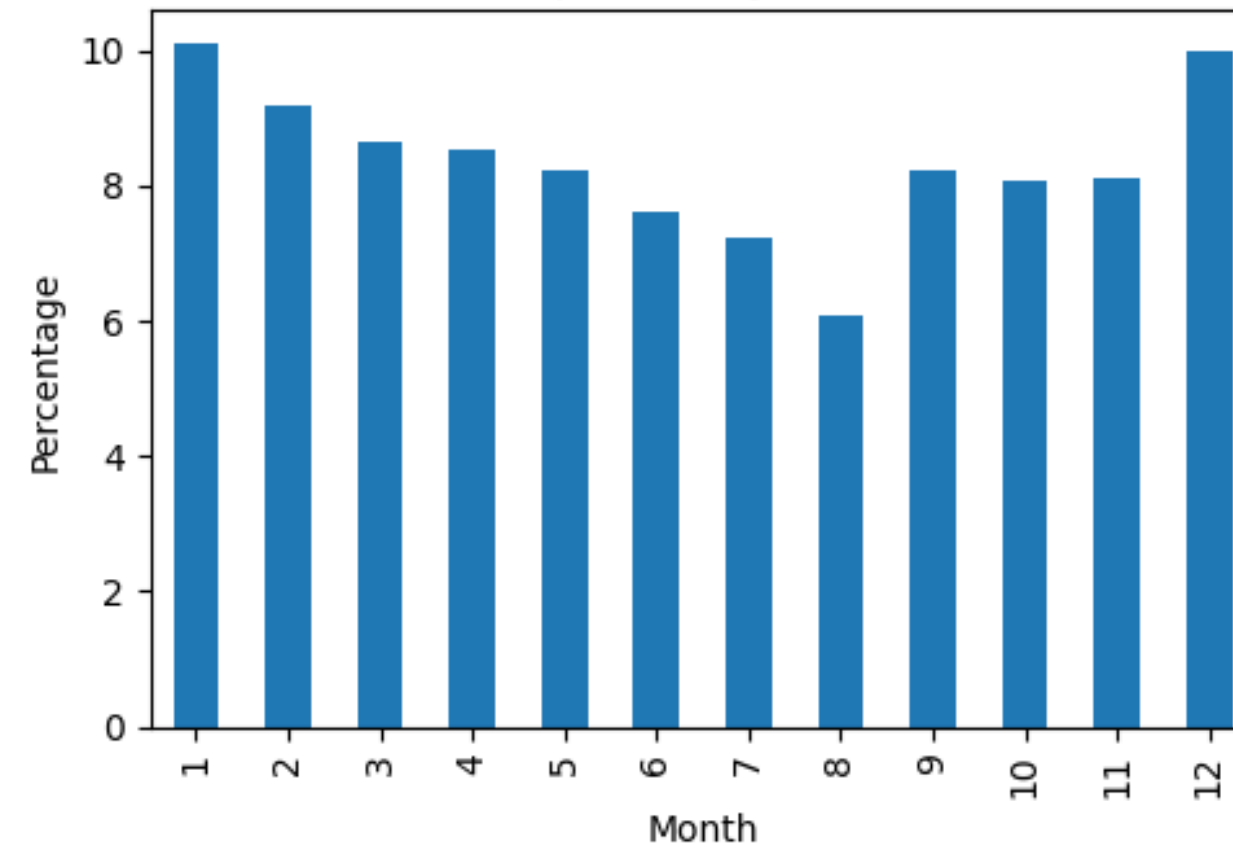
- YEARS AVAILABLE: [2013, 2014, 2016, 2017, 2018, 2019, 2020]
- MONTHS: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
- NUMBER OF INSTALLATION ZONES: 49, VALUES FROM INSTALLATION_ZONE 1 TO INSTALLATION_ZONE 49
- NUMBER OF UNIQUE CONSUMERS: 27632
- CONSUMPTION VALUES RANGING FROM 0 TO 4978 M3
- NUMBER OF WATER CONSUMPTION VALUES EQUAL 0 M3: 61658

PROPORTION OF 0 M3 CONSUMPTION VALUE BY YEAR, MONTH AND CONSUMER_TYPE

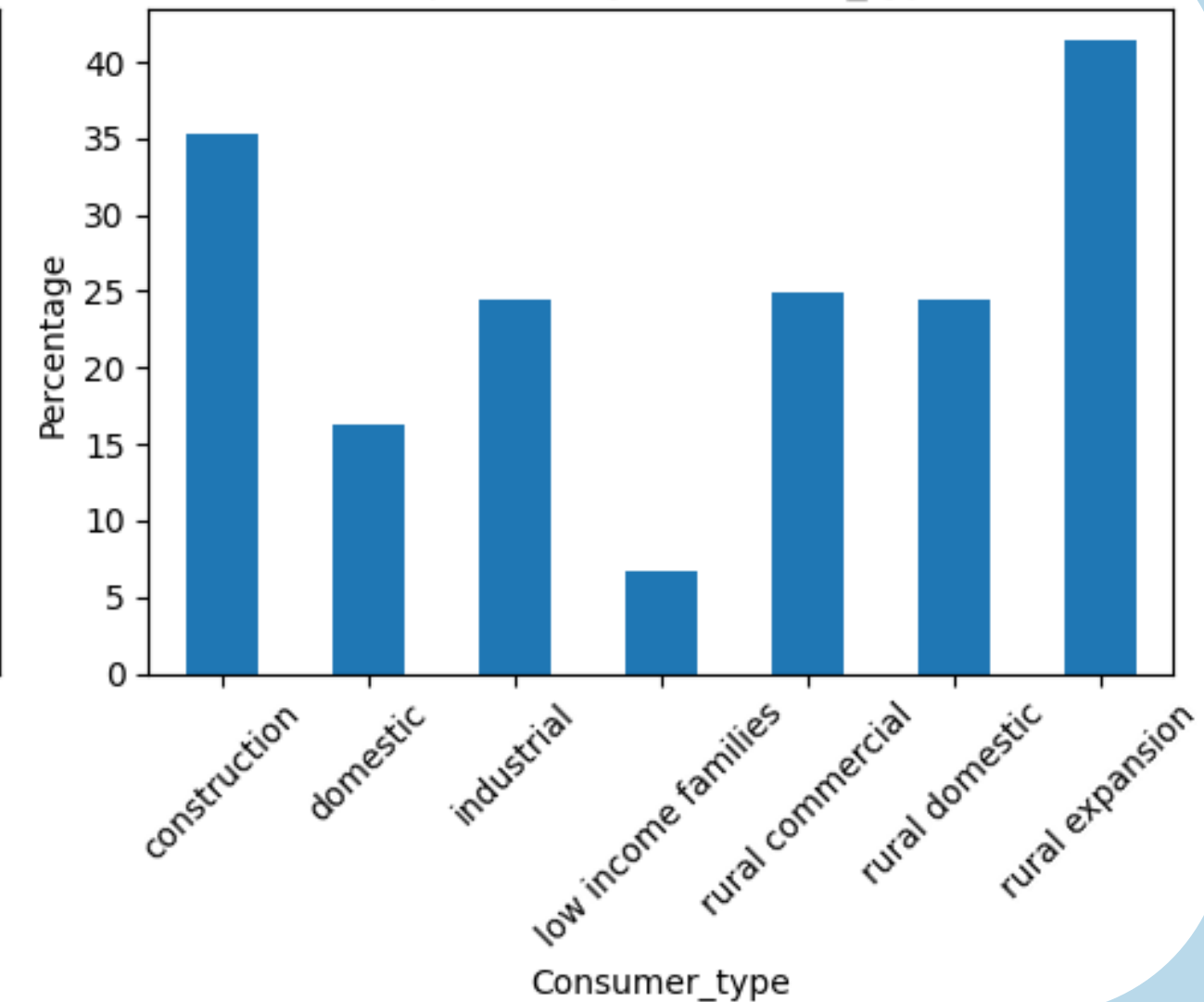
Proportion by Year



Proportion by Month



Proportion by Consumer_type

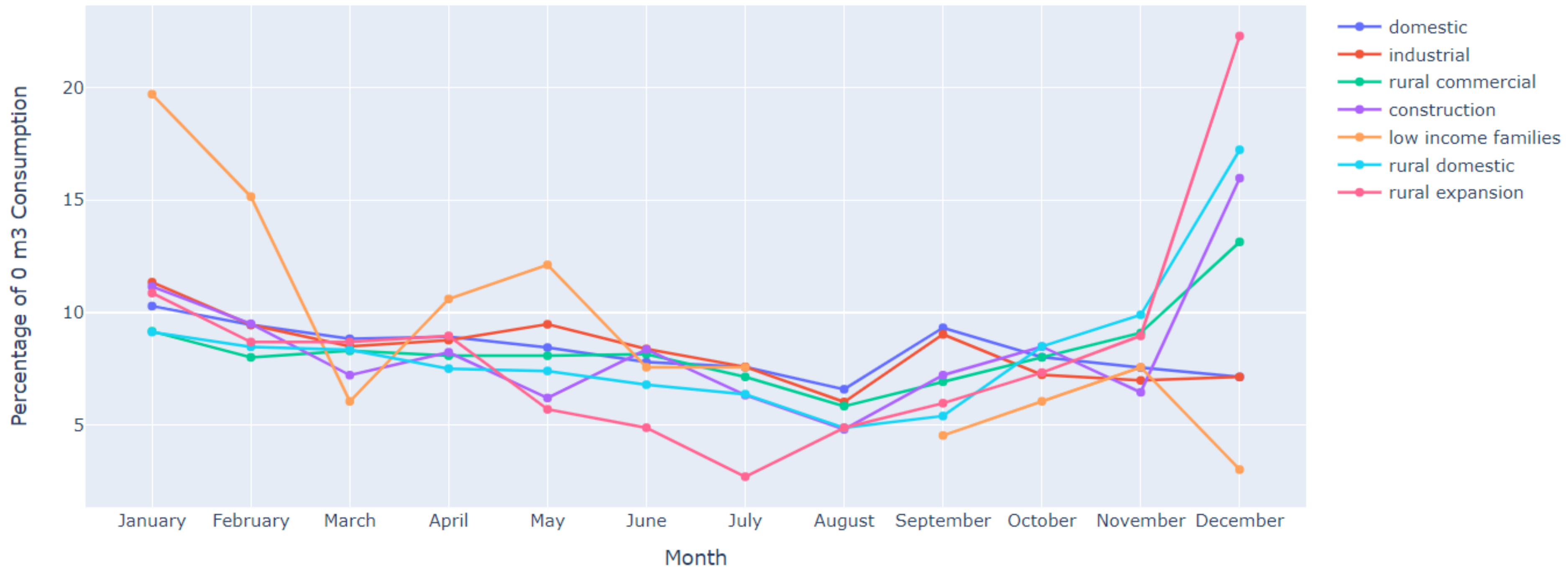




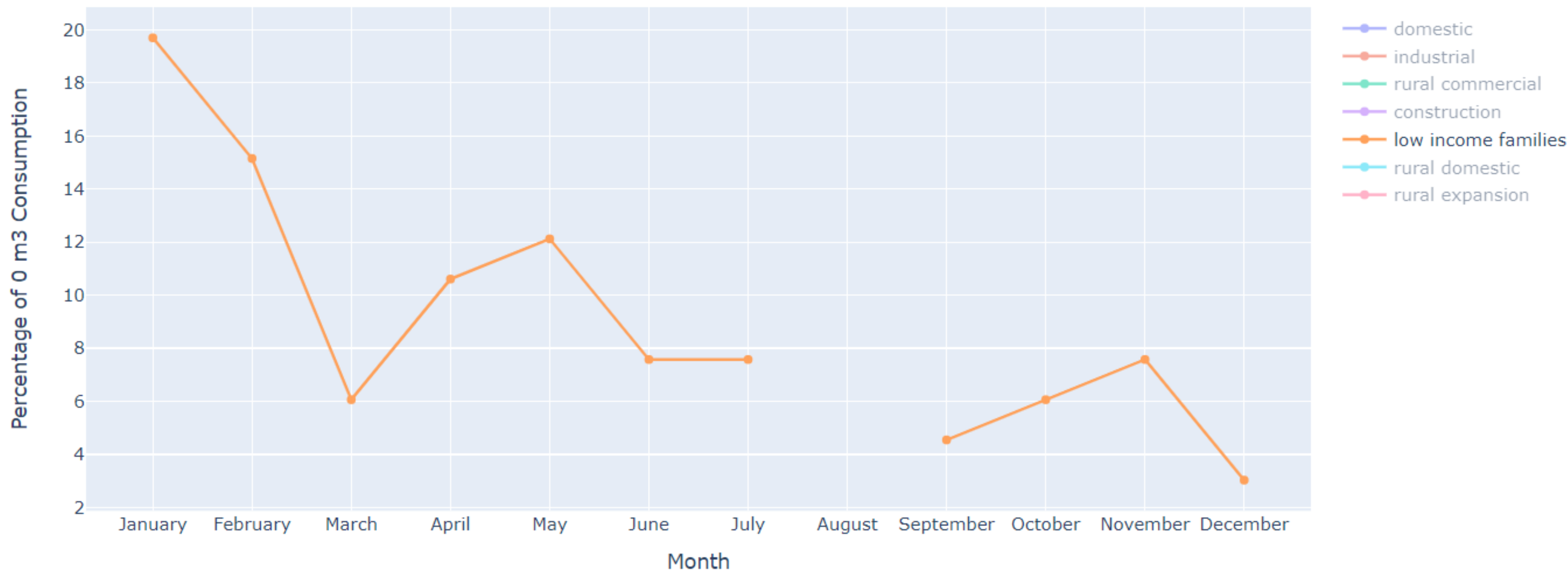
EDA (EXPLORATORY DATA ANALYSIS)

- MISSING CONSUMPTION VALUES
- HANDLING ZERO VALUES
- OUTLIERS
- HEATMAPS AND TEMPORAL TRENDS

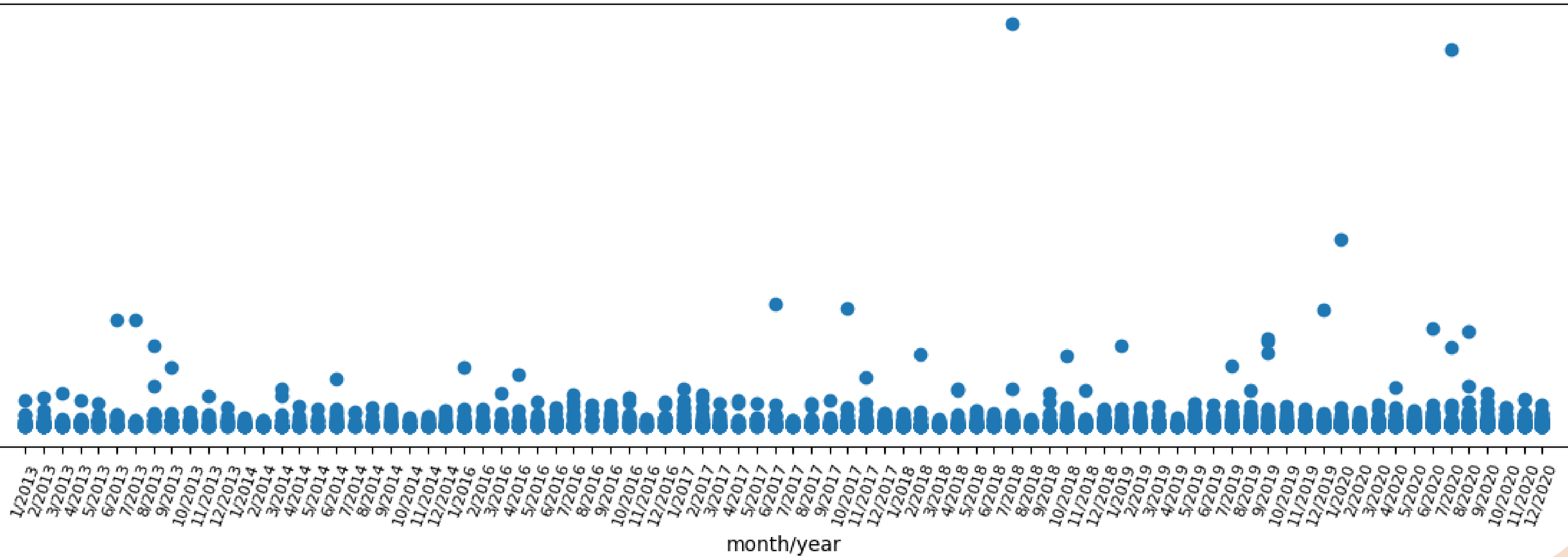
PROPORTION OF 0 M3 CONSUMPTION VALUE BY MONTH AND CONSUMER_TYPE



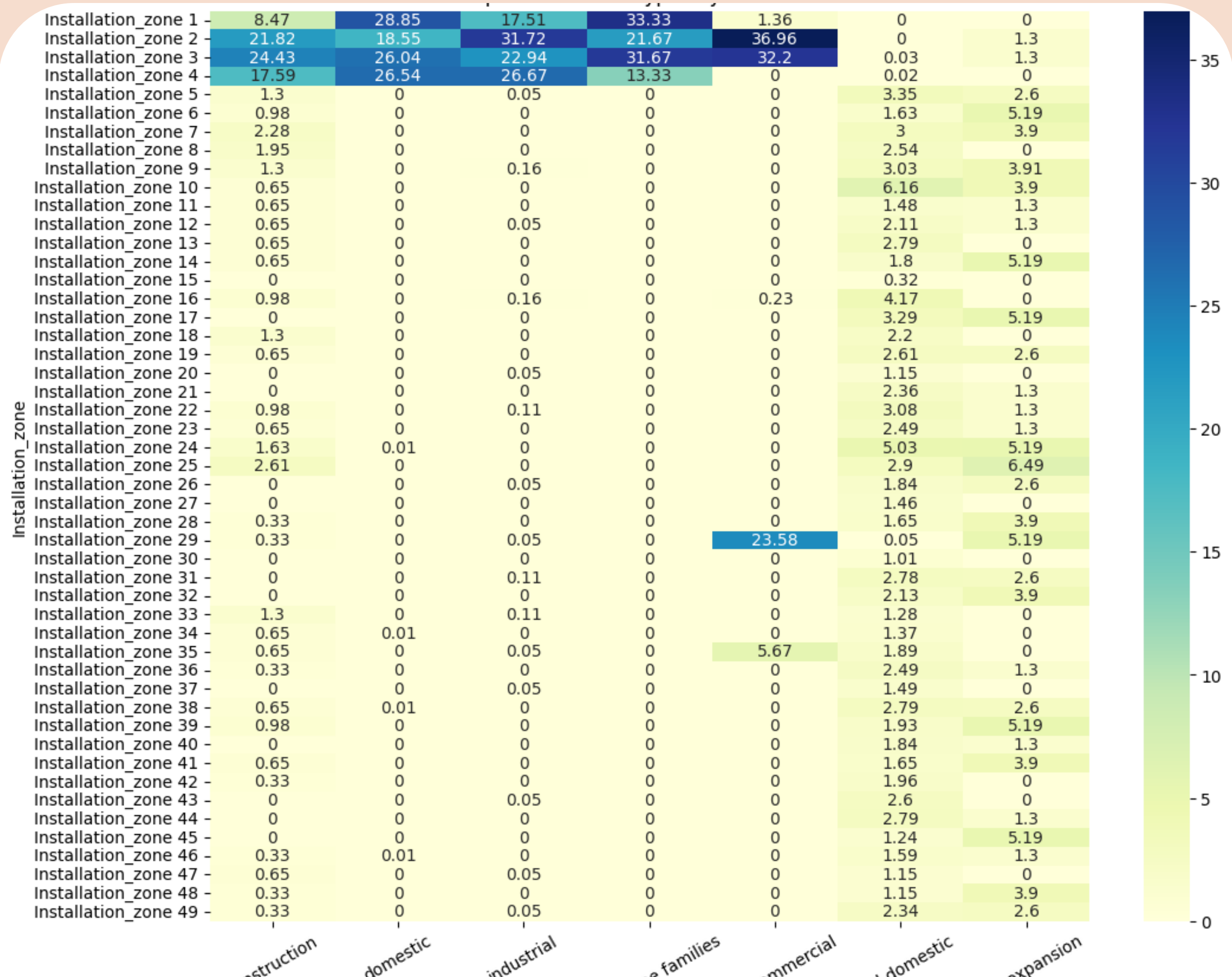
Proportion of 0 m3 Consumption value by Month and Consumer Type



scatter plot construction



HEATMAP OF CONSUMER_TYPES BY INSTALLATION_ZONE



PREPROCESSING

- DATA CLEANING

- NAN PROCESSING

- REMOVING DATA

- IMPUTING DATA

- REMOVING NEGATIVE VALUES

- REMOVING THE OUTLIERS

DATA COMPLETION

- WHEN THERE IS MISSING DATA FOR A PARTICULAR CUSTOMER AND MONTH, THIS ROW OF DATA DOESN'T EXIST IN THE DATASET
- FOR COMPLETING THE DATA WE HAVE ADDED ROWS FOR ALL THE MONTHS, SO THAT WE HAD 600.000 SAMPLE BEFORE AND AROUND 2.000.000 AFTER THE PROCESSING
- NOW WE HAVE TO DEAL WITH ALL THE NEW NAN DATA

DATA REMOVAL AND INPUT

- DATA REMOVAL

- WHEN FOR A YEAR THERE IS ONLY AVAILABLE DATA FOR ONE OR TWO MONTHS, WE DECIDED TO REMOVE THE ENTIRE YEAR FOR THAT CUSTOMER.

- DATA INPUT

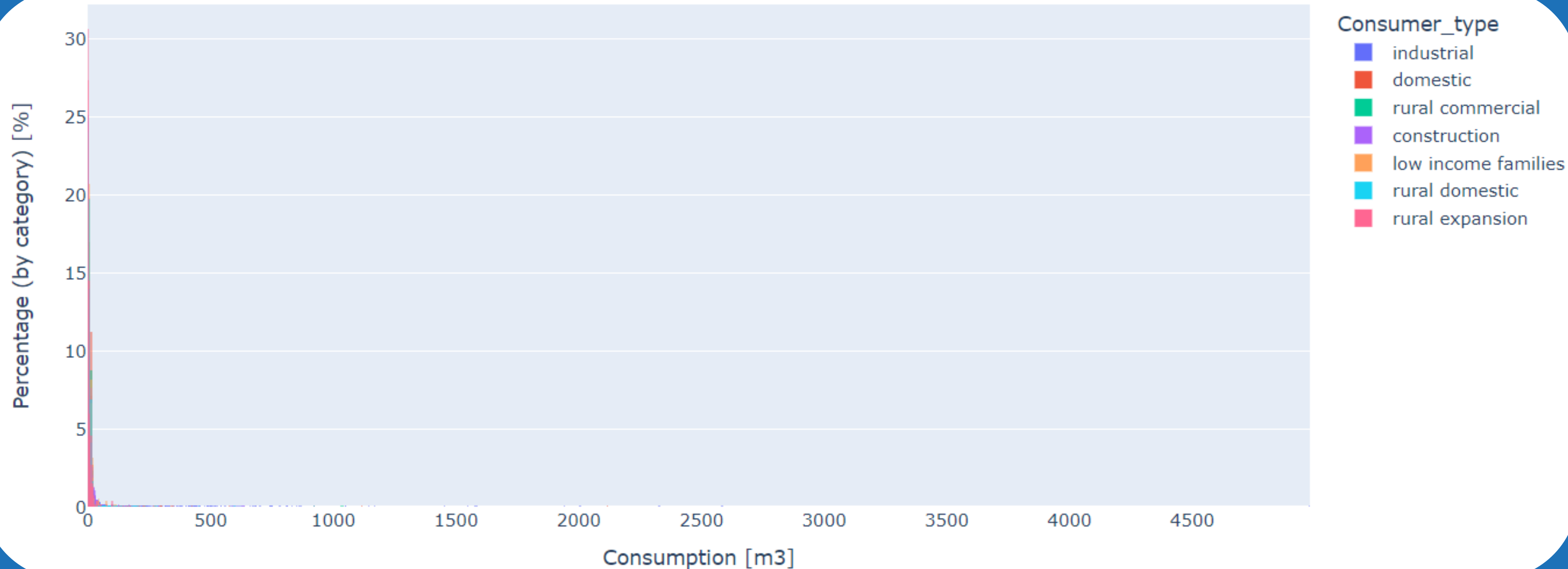
- USERS WITH A NUMBER OF MISSING MONTHS BELOW THE THRESHOLD, IMPUTE THEM WITH THE AVERAGE VALUE OF:
- THE MEAN CONSUMPTION VALUE OF THE CONSUMER FOR A SPECIFIC YEAR
- THE MEAN CONSUMPTION VALUE OF THE CONSUMER_TYPE CATEGORY WICH THE CONSUMER BELONGS TO (FOR A SPECIFIC YEAR)
- THE MONTHLY MEAN CONSUMPTION VALUE OF THE CONSUMER_TYPE CATEGORY WICH THE CONSUMER BELONGS TO (ALL THE YEARS)

MISSING CONSUMPTION VALUES (NaN)

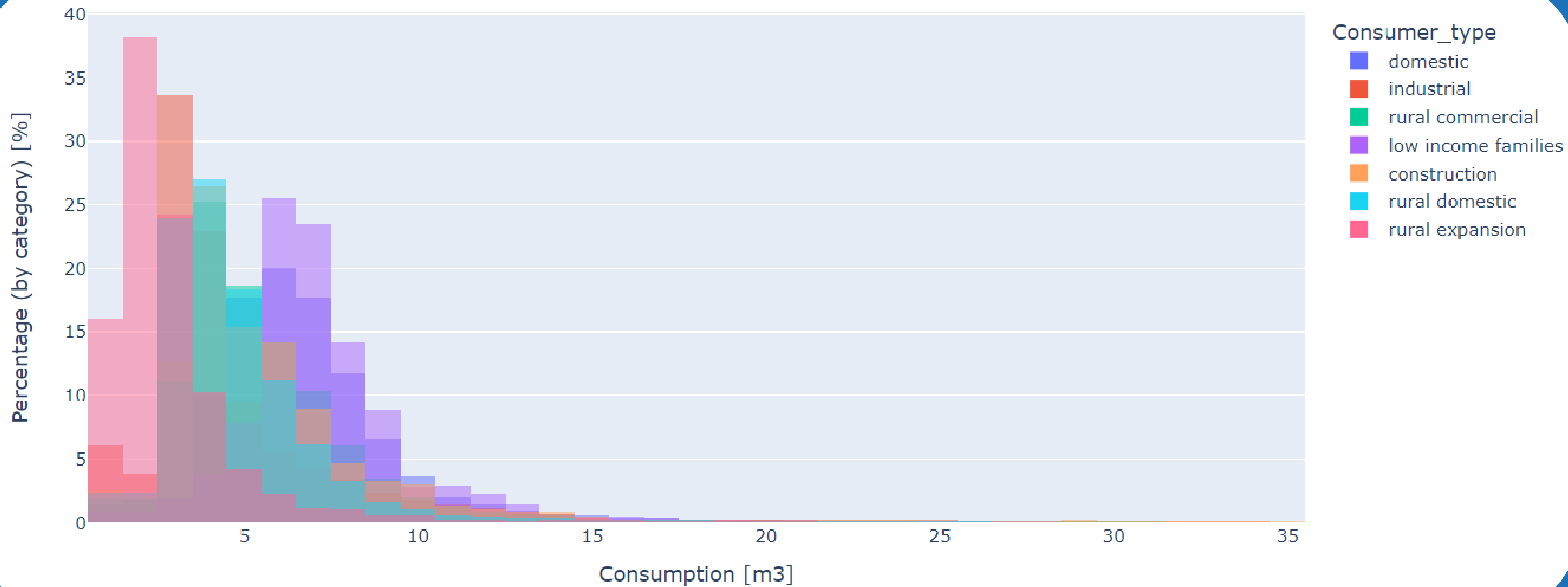
	Year	Month	Consumer_number	Consumer_type	Consumption	Installation_zone	Date
23091	2013	1	AABK96307399687530	NaN	NaN	NaN	NaN
50723	2013	2	AABK96307399687530	NaN	NaN	NaN	NaN
78355	2013	3	AABK96307399687530	NaN	NaN	NaN	NaN
105987	2013	4	AABK96307399687530	NaN	NaN	NaN	NaN
133619	2013	5	AABK96307399687530	NaN	NaN	NaN	NaN
161251	2013	6	AABK96307399687530	NaN	NaN	NaN	NaN
188883	2013	7	AABK96307399687530	NaN	NaN	NaN	NaN
216515	2013	8	AABK96307399687530	NaN	NaN	NaN	NaN
244147	2013	9	AABK96307399687530	NaN	NaN	NaN	NaN
271779	2013	10	AABK96307399687530	NaN	NaN	NaN	NaN
299411	2013	11	AABK96307399687530	NaN	NaN	NaN	NaN
327043	2013	12	AABK96307399687530	NaN	NaN	NaN	NaN
354675	2014	1	AABK96307399687530	NaN	NaN	NaN	NaN
382308	2014	2	AABK96307399687530	NaN	NaN	NaN	NaN
409942	2014	3	AABK96307399687530	NaN	NaN	NaN	NaN
437576	2014	4	AABK96307399687530	NaN	NaN	NaN	NaN

	Year	Month	Consumer_number	Consumer_type	Consumption	Installation_zone	Date	Consumer_Avg	Group_Avg	Group_Monthly_Avg
44475	2018	1	AABK96307399687530	domestic	6.036	Installation_zone 1	2018-01-01	5.667	6.228	6.214
108049	2018	2	AABK96307399687530	domestic	5.998	Installation_zone 1	2018-02-01	5.667	6.228	6.098
171662	2018	3	AABK96307399687530	domestic	5.000	Installation_zone 1	2018-03-01	5.667	6.228	5.908
235261	2018	4	AABK96307399687530	domestic	6.063	Installation_zone 1	2018-04-01	5.667	6.228	6.294
298829	2018	5	AABK96307399687530	domestic	6.176	Installation_zone 1	2018-05-01	5.667	6.228	6.634
362381	2018	6	AABK96307399687530	domestic	6.202	Installation_zone 1	2018-06-01	5.667	6.228	6.710
425900	2018	7	AABK96307399687530	domestic	6.188	Installation_zone 1	2018-07-01	5.667	6.228	6.668
489465	2018	8	AABK96307399687530	domestic	5.000	Installation_zone 1	2018-08-01	5.667	6.228	6.919
552824	2018	9	AABK96307399687530	domestic	6.191	Installation_zone 1	2018-09-01	5.667	6.228	6.678
616405	2018	10	AABK96307399687530	domestic	6.089	Installation_zone 1	2018-10-01	5.667	6.228	6.373
679989	2018	11	AABK96307399687530	domestic	6.131	Installation_zone 1	2018-11-01	5.667	6.228	6.499
743589	2018	12	AABK96307399687530	domestic	7.000	Installation_zone 1	2018-12-01	5.667	6.228	6.624
52973	2019	1	AABK96307399687530	domestic	6.608	Installation_zone 1	2019-01-01	7.250	6.359	6.214
116560	2019	2	AABK96307399687530	domestic	6.569	Installation_zone 1	2019-02-01	7.250	6.359	6.098
180171	2019	3	AABK96307399687530	domestic	7.000	Installation_zone 1	2019-03-01	7.250	6.359	5.908
243767	2019	4	AABK96307399687530	domestic	6.634	Installation_zone 1	2019-04-01	7.250	6.359	6.294
307334	2019	5	AABK96307399687530	domestic	6.748	Installation_zone 1	2019-05-01	7.250	6.359	6.634
370896	2019	6	AABK96307399687530	domestic	6.773	Installation_zone 1	2019-06-01	7.250	6.359	6.710
434415	2019	7	AABK96307399687530	domestic	6.759	Installation_zone 1	2019-07-01	7.250	6.359	6.668
497967	2019	8	AABK96307399687530	domestic	6.843	Installation_zone 1	2019-08-01	7.250	6.359	6.919

CONSUMPTION DISTRIBUTION FOR EACH CONSUMER_TYPE BEFORE REMOVING OUTLIERS

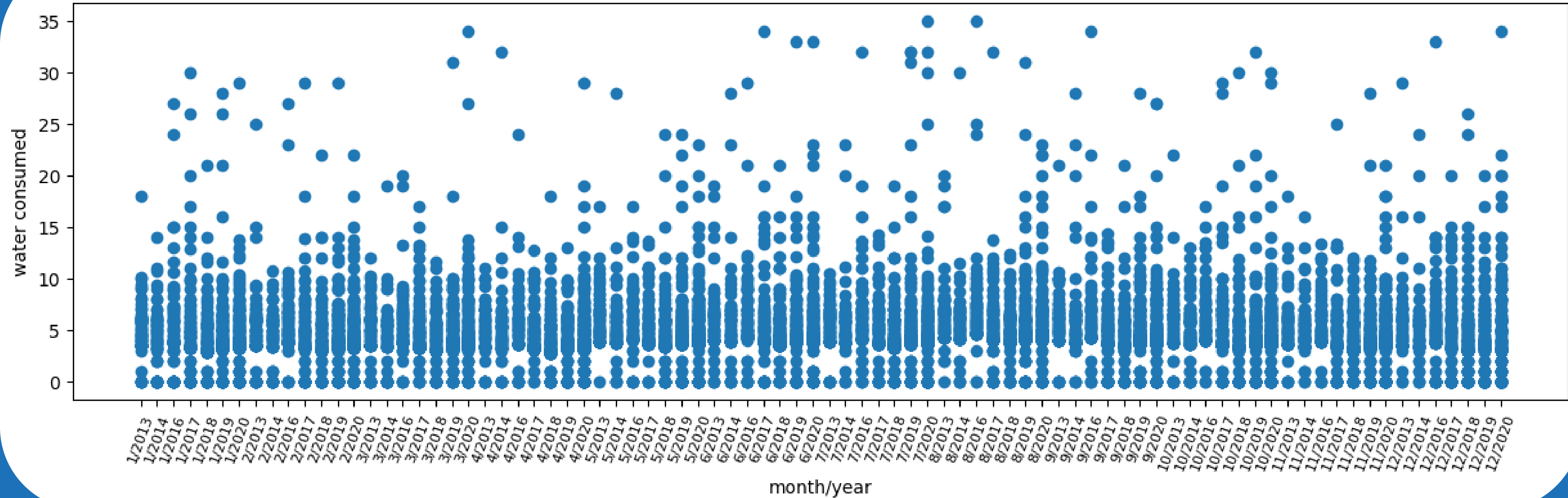


AFTER



AFTER

scatter plot construction



DATA TRANSFORMATION

- TRAIN/TEST SPLIT

- FEATURE ENCODING

- CYCLICAL FEATURES

- CATEGORICAL FEATURES

- FEATURE SCALING

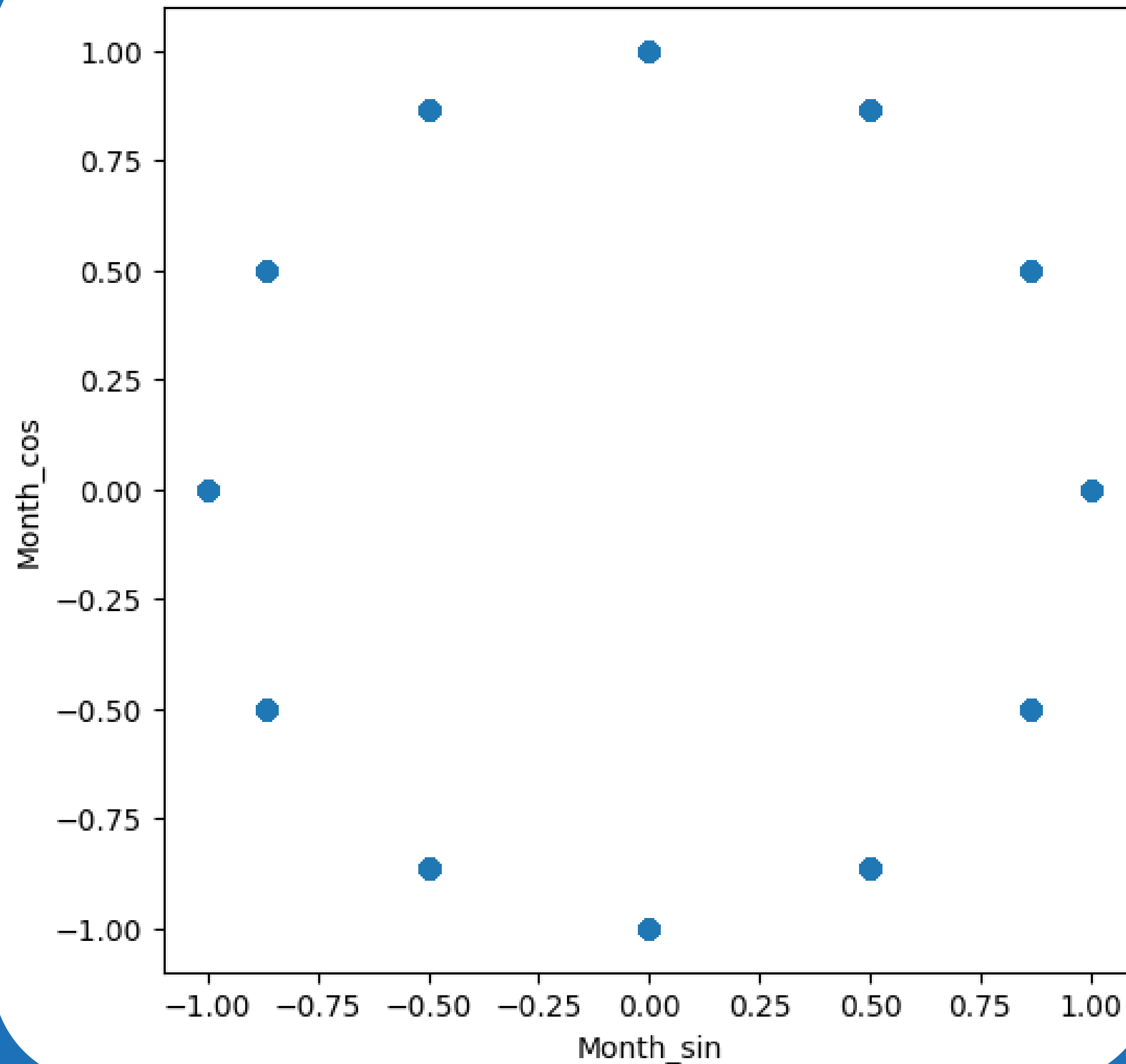
- STANDARDIZATION

- MIN-MAX SCALING (NORMALIZATION)

- ROBUST SCALING

- POWER TRANSFORMATION

Parametric plot of the cyclical feature "Month"



CATEGORICAL FEATURES

- ONE-HOT ENCODING
- LABEL ENCODING
- FREQUENCY ENCODING
- HASHING ENCODING
- TARGET ENCODING

METHOD USED (TARGET ENCODING)

USEFUL FOR:

- HIGH CARDINALITY FEATURES: FEATURES WITH A LARGE NUMBER OF UNIQUE VALUES.
- IMBALANCED CATEGORIES
- PREVENTS THE GENERATION OF SPARSE FEATURES AS IN ONE-HOT ENCODING.

FEATURE SCALING

- STANDARDIZATION
- MIN-MAX SCALING (NORMALIZATION)
- ROBUST SCALING
- POWER TRANSFORMATION

ADJUSTMENT AND PREPARATION OF THE MODEL

PROJECT LINK:

[HTTPS://GITHUB.COM/RCHATRU/ML4DS](https://github.com/rchatru/ml4ds)



THANK
YOU! <3