BHT Berliner Hochschule für Technik

Geo-Clustering

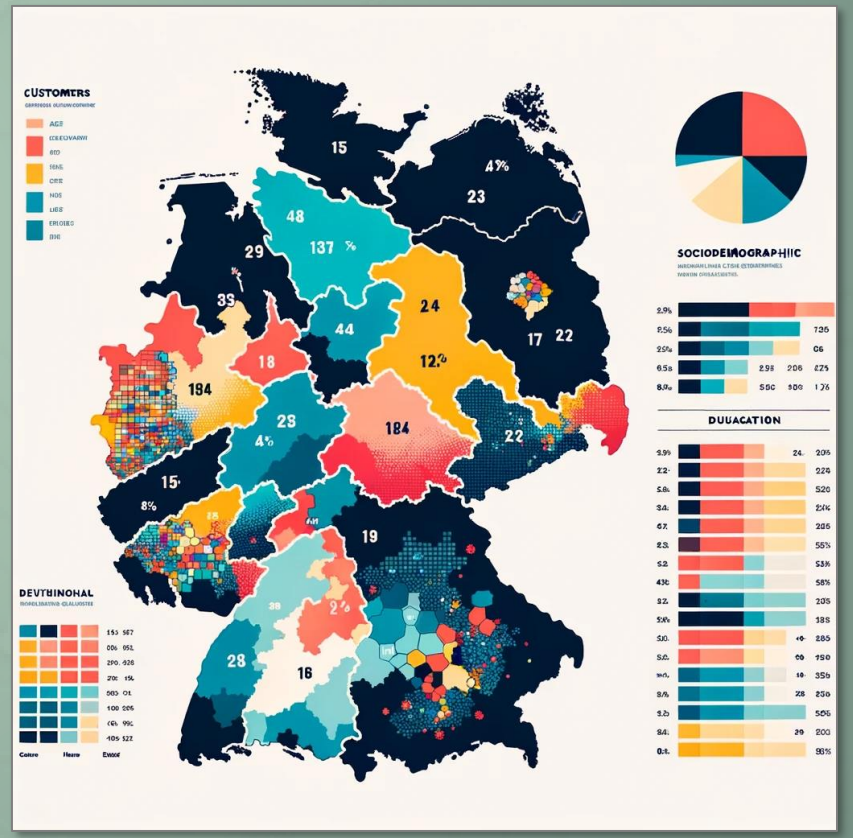
Advanced Software-Engineering

Dr. Harald Stein, Prof. Dr.-Ing. Stefan Edlich

Jan 2024

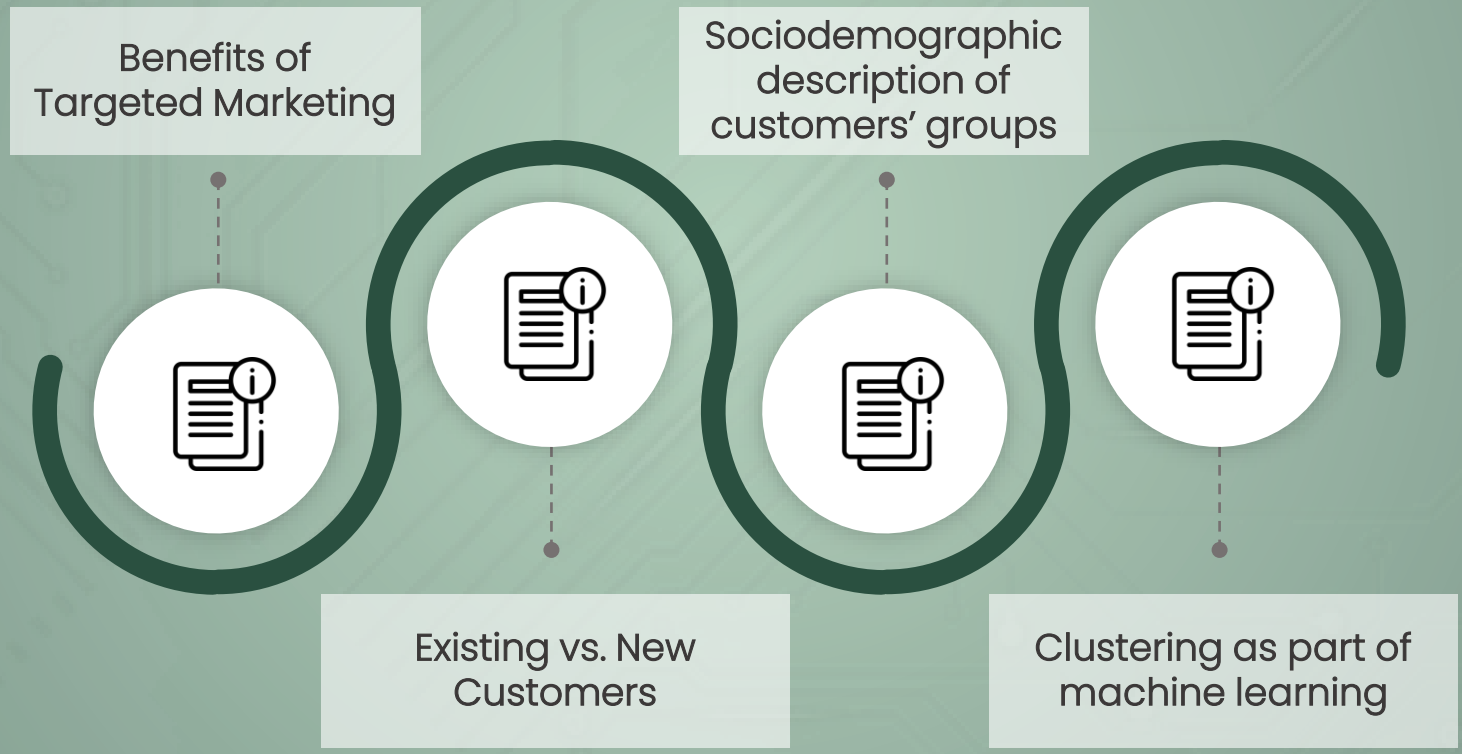
Agenda

- Introduction
- Data Basis for sociodemographic Geoclustering
- Feature selection with PCA and Matrix rotation
- Clustering
- Effect Analysis Tables
- Geovisualization



Introduction

Targeted marketing for existing and new customers in particular sociodemographic segments and geolocations



Benefits of Targeted Marketing

... enables businesses to direct resources towards specific segment of audience most likely to be interested in their product/service, improving conversion rates, return on investment

Increased Conversion Rates

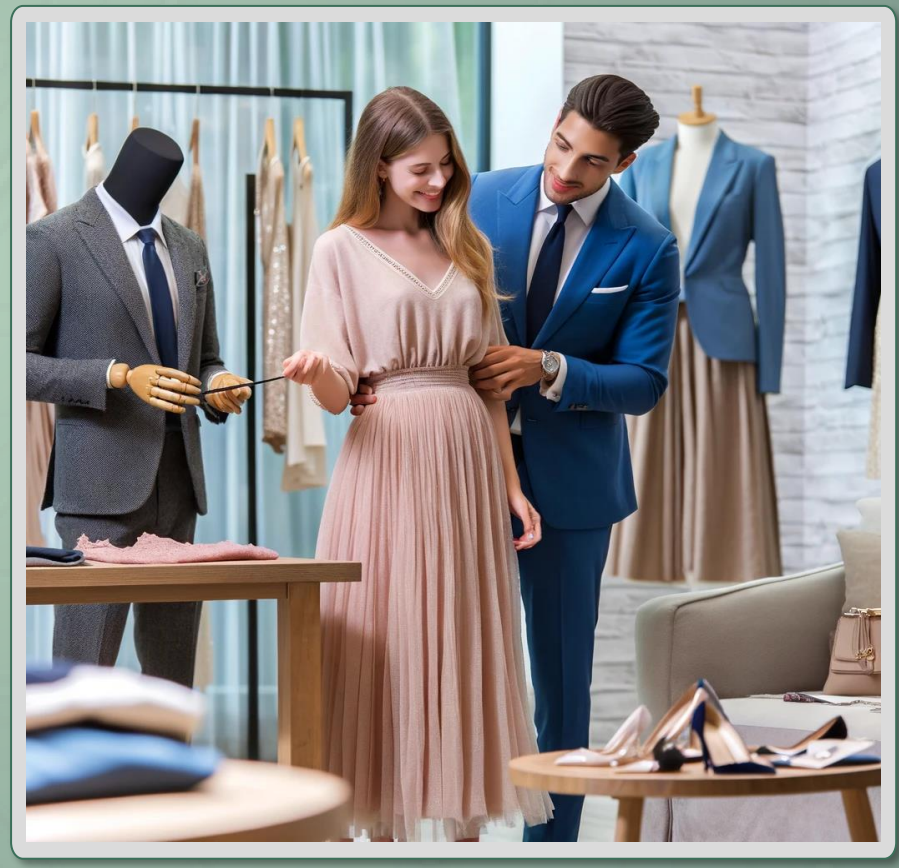
- Tailoring marketing strategies to a specific audience increases the likelihood of engagement and conversion.
- Personalization makes consumers more likely to act because the content resonates with their specific needs and preferences.

Cost Efficiency

- By focusing resources on a defined group, businesses can allocate their budgets more effectively.
- Reduces the waste of marketing dollars on broad, undefined audiences.

Enhanced Customer Relationships

- Targeted marketing fosters a deeper connection with consumers by addressing their unique challenges and desires.
- Builds loyalty and trust by showing that the business understands and values the customer's individual journey.



Existing vs. New Customers

Focus: Sociodemographic segments of existing customers

Focus	Existing Customers	New Customers
Understanding the Audience	▪ Analyse historical data understanding segments, preferences	▪ Use market research, predictions to identify potential needs.
Communication Strategy	▪ Focus on customers or segments with high turnover, loyalty	▪ Emphasize awareness and brand introduction.
Customized Offers	▪ Tailor promotions based on past behavior and preferences.	▪ Provide introductory offers to attract trial and engagement.
Feedback Loop	▪ Gather and act on feedback for continuous improvement.	▪ Use initial interaction to understand preferences, expectations.

Sociodemographic description

... of customers' groups

Population in house

Status of Location

Tendency of having a photovoltaic system

Tendency of having children

Main purchase criterion: low price

Population ratio: university degree

Interest: football

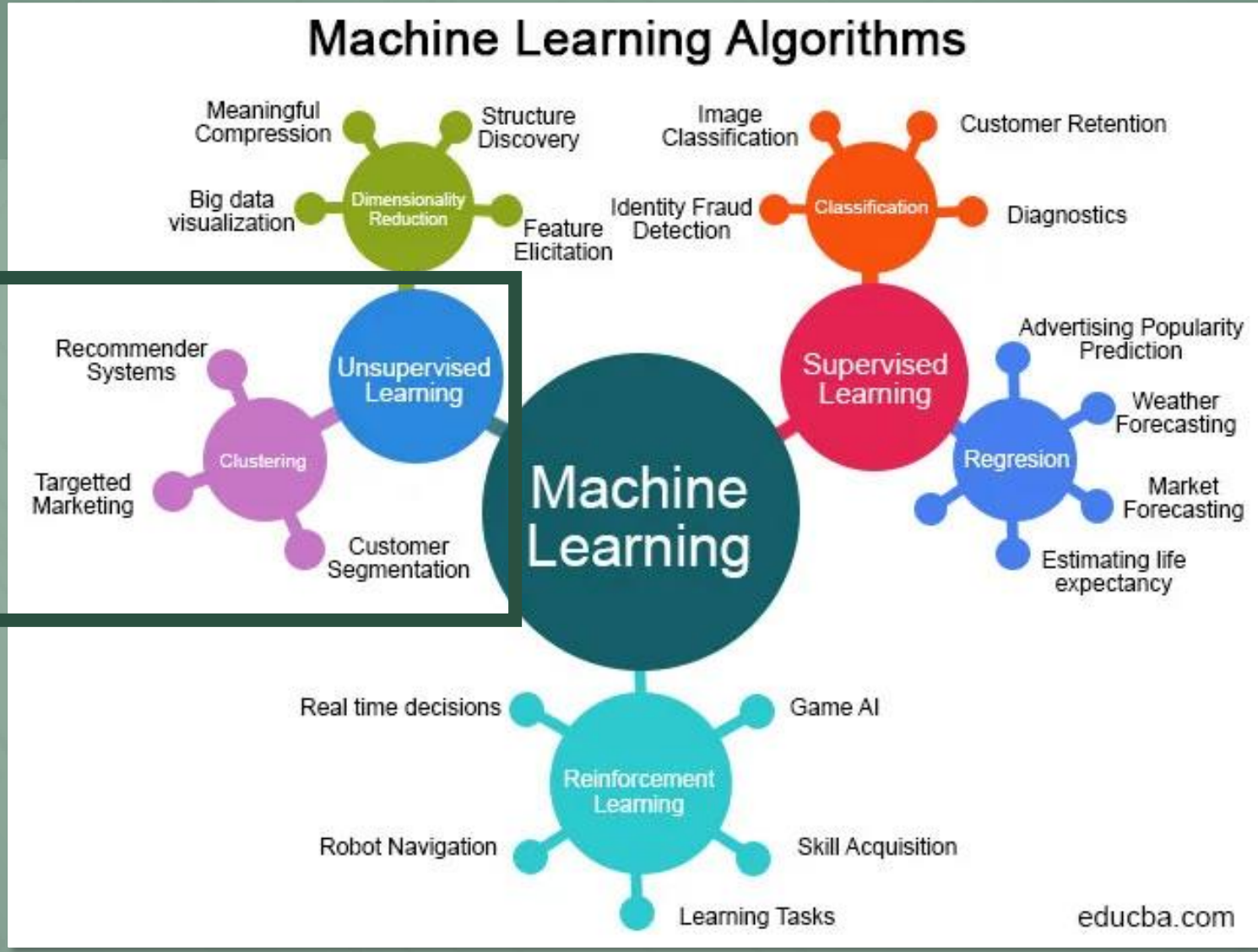
Interest: online shopping

Interest: travel

Main purchase criterion: quality

Clustering/Segmentation as part of Machine Learning

Focus



Data Basis for sociodemographic Geoclustering

... from joining sales events, geodata and sociodemographic data and anonymizing it by aggregation

What is Sociodemographic Geoclustering?

… method combining geographic data, demographic characteristics to identify and segment similar populations within specific areas.

Key Components

- Geographic Information:
  Physical locations defined by boundaries, natural markers.
- Sociodemographic Analysis:
  Age, income, education, ethnicity, other social variables.

Data Synthesis, Integrates various data sources:

- census data
- consumer behavior
- geographic information systems (GIS)

Example: Coffee culture in different areas

https://www.tip-berlin.de/kultur/tip-berlin-karten-galerie/

Why Use Sales Events with Address?

Marking locations with sales events is basis for geo-located marketing activities

Granular Customer Insights

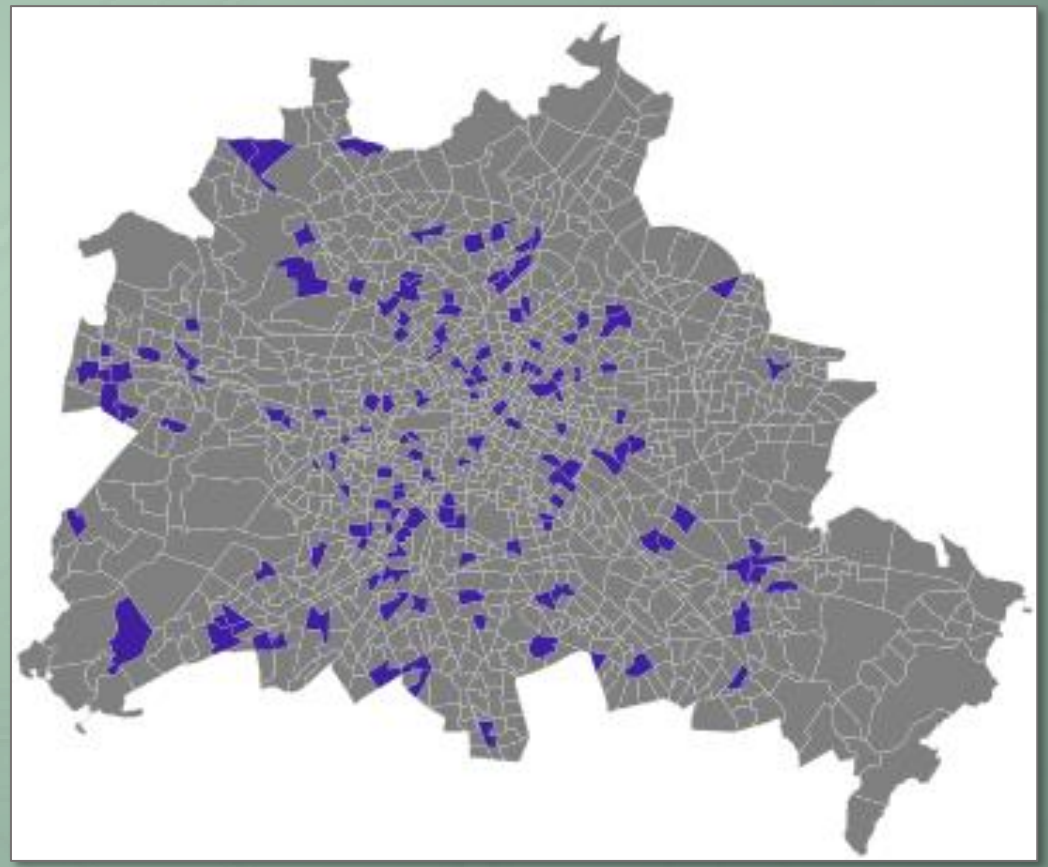
Addresses linked with sales events paint detailed picture of consumer behavior within specific locales.

Enhanced Targeting

- Utilizing geographic data enables precise targeting for marketing campaigns and resource allocation.
- Evaluates sales trends and the effectiveness of promotional activities by region.

Basis for further analysis

- Segmentation of existing customers
- Prediction of areas with new customers



Geocoding

… is the process of converting street addresses into geographic coordinates, i.e. Polygons (latitude and longitude) in order to place them on a map.

Use

- Adds valuable layer to datasets, enabling correlation with geographic factors for richer insights.
- Enables the representation of address-based data on maps for better visualization and spatial analysis.
- Facilitates strategic placement of services and resources by identifying high-demand areas.

John F. Kennedy Platz, Schöneberg

… is coded to:

POLYGON ((13.36209 52.489335, 13.359698 52.488577, 13.36063 52.488147, 13.360839 52.488022, 13.35776 52.486818, 13.357598 52.486711, 13.356968 52.486964, 13.356757 52.486666, 13.356496 52.486365, 13.356189 52.486183, 13.356034 52.48609, 13.355834 52.4863, 13.354686 52.487503, 13.353676 52.488723, 13.353436 52.489065, 13.35319 52.489915, 13.353888 52.489966, 13.354046 52.489942, 13.354979 52.490056, 13.356376 52.490207, 13.35914 52.490487, 13.359261 52.490501, 13.35942 52.490562, 13.359691 52.490557, 13.360011 52.490554, 13.3603 52.49048, 13.360392 52.490455, 13.360789 52.490246, 13.360842 52.490136, 13.36209 52.489335))

…additionally short-cut geocode is given for convenience:

10356

Merging sales events with sociodemographic data

… provides multifaceted understanding of consumer behavior.

Sales events

Index	CustomerID	PackageID	PackageName	EventDate	Geocode
0	C00001	P001	City Break	2023-08-25	11132
1	C00002	P001	Safari Expedition	2023-11-04	10175
2	C00003	P003	Safari Expedition	2024-10-14	10752
3	C00004	P002	City Break	2023-10-17	10507
4	C00005	P002	Cruise Getaway	2024-09-09	10951

Sociodemographic data

Index	Geocode	Population in the house	House size: number of flats	Status of Location	Tendency of having a photovoltaic system
0	10001	16.7805	1.77591	1.95978	2.97616
1	10002	3.15143	0.635435	2.32054	1.6231
2	10006	7.99221	0.0784368	1.07671	2.54852
3	10007	38.9127	2.90708	2.23876	2.55171
4	10008	33.5625	0.681475	2.51171	3.37949

Merge

Index	CustomerID	PackageID	PackageName	EventDate	Geocode	Population in the house	House size: number of flats	Status of Location
0	C00094	P002	Mountain Adventure	2024-11-26	10013	11.6534	1.58083	0.900039
1	C00348	P003	Mountain Adventure	2024-07-16	10013	11.6534	1.58083	0.900039
2	C00549	P004	Safari Expedition	2024-05-01	10013	11.6534	1.58083	0.900039
3	C01077	P004	Mountain Adventure	2024-06-08	10013	11.6534	1.58083	0.900039
4	C01206	P001	Cruise Getaway	2024-12-22	10013	11.6534	1.58083	0.900039

Be careful of legal pitfall:

- Names, addresses are particularly regulated under law of protection of data privacy
- EU: Datenschutz-Grundverordnung (DSGVO) / General Data Protection Regulation (GDRP)
- Personal data is only allowed to be used if person allows it explicitly for particular cause and time interval
- Even coded / pseudonymized data falls under DSGVO/GDRP, if inference to person is possible

Bypass personal data protection through aggregation

Aggregating data is a way to anonymize personal details, to align with data protection laws while still yielding valuable insights

Group by field

Index	CustomerID	PackageID	PackageName	EventDate	Geocode	Population in the house	House size: number of flats	Status of Location
0	C00094	P002	Mountain Adventure	2024-11-26	10013	11.6534	1.58083	0.900039
1	C00348	P003	Mountain Adventure	2024-07-16	10013	11.6534	1.58083	0.900039
2	C00549	P004	Safari Expedition	2024-05-01	10013	11.6534	1.58083	0.900039
3	C01077	P004	Mountain Adventure	2024-06-08	10013	11.6534	1.58083	0.900039
4	C01206	P001	Cruise Getaway	2024-12-22	10013	11.6534	1.58083	0.900039

Group by field

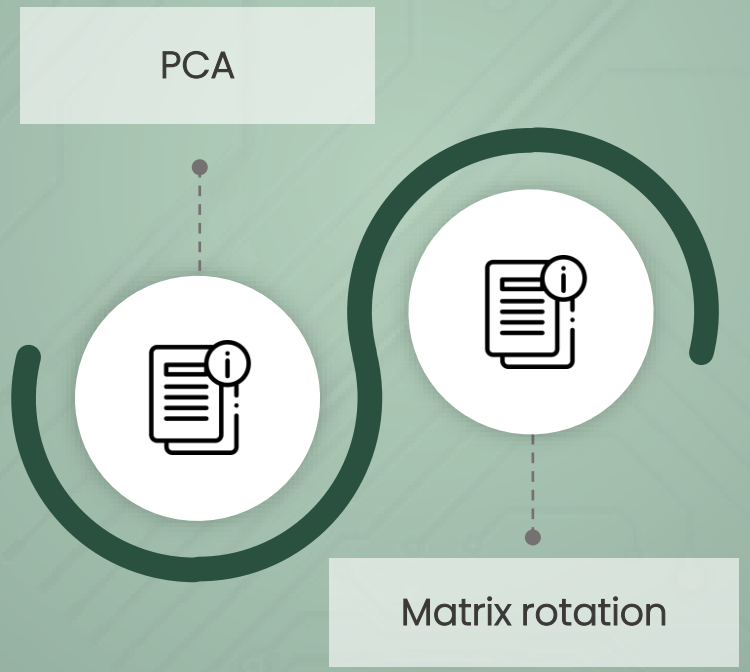
averaged fields

Index	Geocode	Population in the house	House size: number of flats	Status of Location	Tendency of having a photovoltaic system	Tendency of having children
0	10013	1.6534	1.58083	0.900039	4.01617	5.09912
1	10015	2.2082	2.1618	3.31715	1.70406	4.95647
2	10026	3.5425	1.53105	4.3079	2.15104	3.39017
3	10027	0.9565	1.86862	4.17824	3.8672	4.40419
4	10032	.26364	1.1363	3.64194	5.30521	2.59223

(... all sociodemographic fields)

Feature selection with PCA and Matrix rotation

... involves reducing the dimensionality of data by identifying principal features that are mostly independent and maximize variance which is considered as measure of information



Principal Component Analysis (PCA)

PCA is a statistical technique used to simplify the complexity in high-dimensional data while retaining trends and patterns.

Dimensionality Reduction

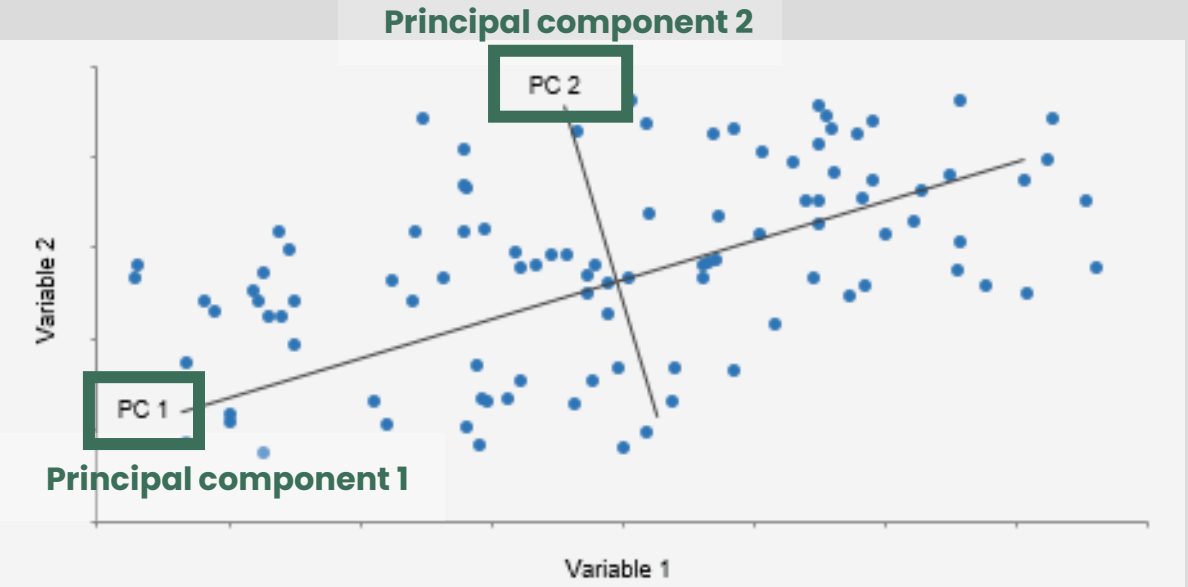
It reduces number of variables by transforming original set into new set of uncorrelated features called principal components.

Enhancing Visualization

By reducing dimensions, PCA enables the visualization of complex datasets in two or three dimensions.

Improving Model Performance

Used in feature selection to enhance model performance by eliminating multicollinearity and noise.



Matrix rotation

… is transformation technique applied to principal components loading matrix to select most appropriate original features

Here: Varimax (orthogonal rotation to maximize variance)

Facilitating Interpretation

Rotation makes it easier to interpret the underlying factors by seeking to maximize high loadings and minimize low loadings.

Enhancing Feature Clarity

Helps in distinguishing between features by clarifying which variables associated with which underlying components or factors.

Application in Feature Selection

Utilized to refine feature selection by identifying the most relevant features contributing to the principal components.

Selected features

- Population in the house
- Interest: travel
- Interest: online shopping
- Interest: travel

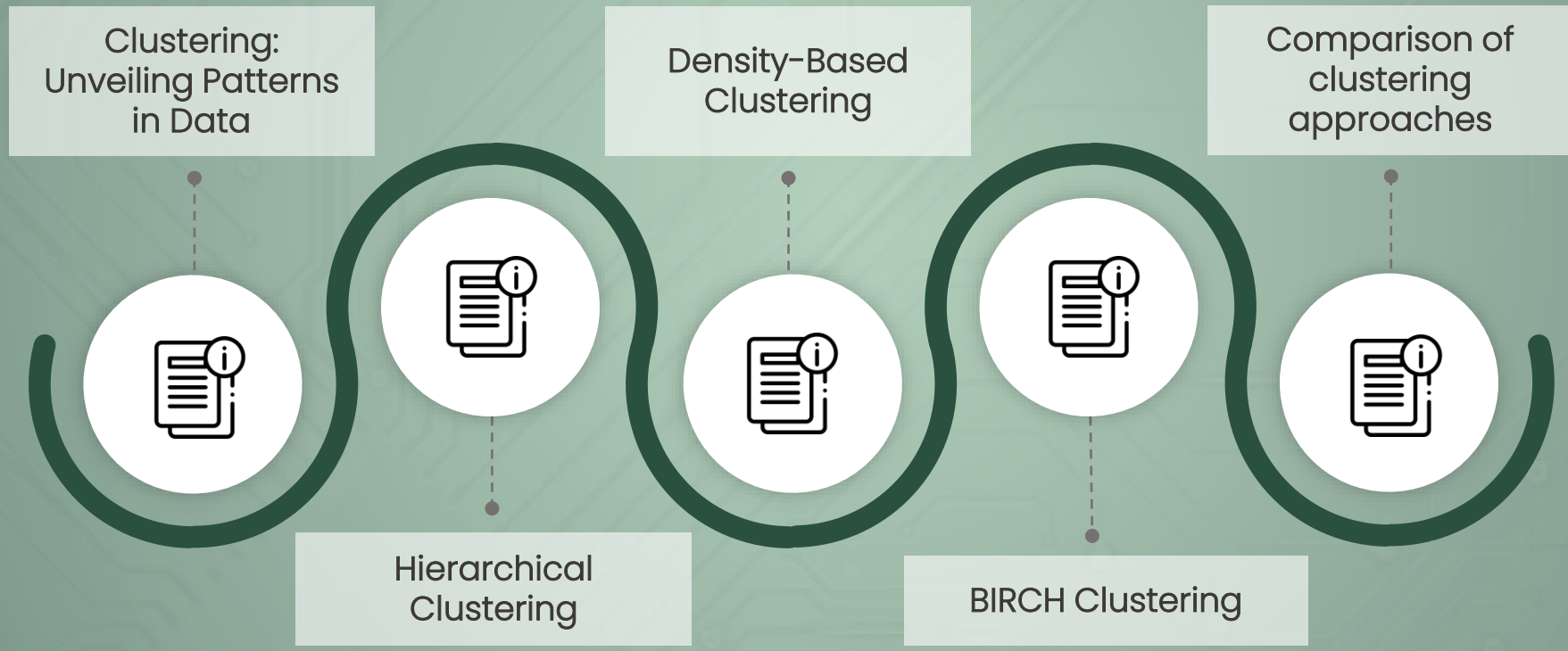
Principal components

Table 4. Varimax rotation PCA loading matrix

	PC1	PC2	PC3	F4
EC	*0.89*	0.13	0.10	-0.14
pH	0.31	*0.67*	-0.14	-0.28
Ca	*0.81*	-0.16	0.23	-0.04
Mg	*0.88*	-0.10	0.32	0.13
Na	*0.88*	-0.21	0.00	0.13
K	*0.58*	-0.11	-0.45	-0.20
Cl	*0.91*	-0.21	0.02	-0.01
SO_4	0.36	*-0.58*	-0.11	0.22
HCO_3	0.41	*0.70*	0.04	-0.12
Pb	0.21	*0.56*	0.38	0.13
Fe	-0.24	0.18	*0.59*	0.50
Zn	0.27	*0.61*	-0.41	0.30
Cu	0.15	0.11	-0.44	*0.77*
Eigenvalue	4.76	2.17	1.26	1.19
Variance (%)	36.60	16.68	9.72	9.15
Cumulative variance %	36.60	53.27	62.99	72.14

Clustering

... is a method of unsupervised learning that groups data points into subsets or clusters based on similarity.



Clustering: Unveiling Patterns in Data

Clustering is direction of unsupervised learning where set of objects or data points are grouped together based on their similarities

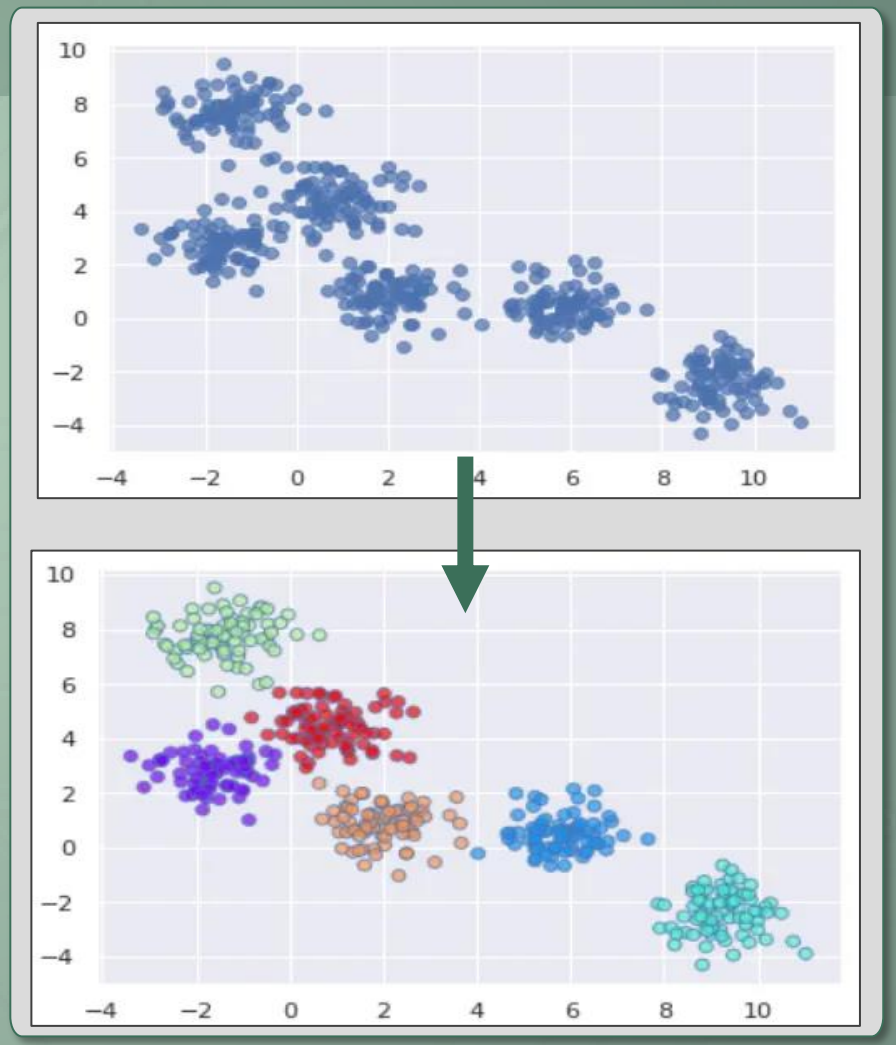
… or to put observations into meaningful groups where members of cluster are more like each other than those in other clusters.

Common Clustering Techniques

- K-MEANS:
  partitions dataset into K distinct, non-overlapping subgroups around centroids, minimizing variance within each cluster, maximizing the variance between different clusters
- Hierarchical clustering (divisive approach):
  builds hierarchy of clusters by recursively dividing large cluster into smaller ones
- DBScan:
  identifies clusters as high-density areas separated by low density areas

Some applications Across Fields

- Market segmentation
- Object detection in images
- Document clustering
- Anomaly detection.



Hierarchical Clustering

… is method of cluster analysis which seeks to build hierarchy of clusters through successive iterations.

- Divisive (top-down) approach:
  starts with all points in one cluster, divides into smaller clusters.
- Agglomerative (bottom-up) approach:
  starts with each data point as single cluster, merges them step by step

Visualizing with Dendrograms

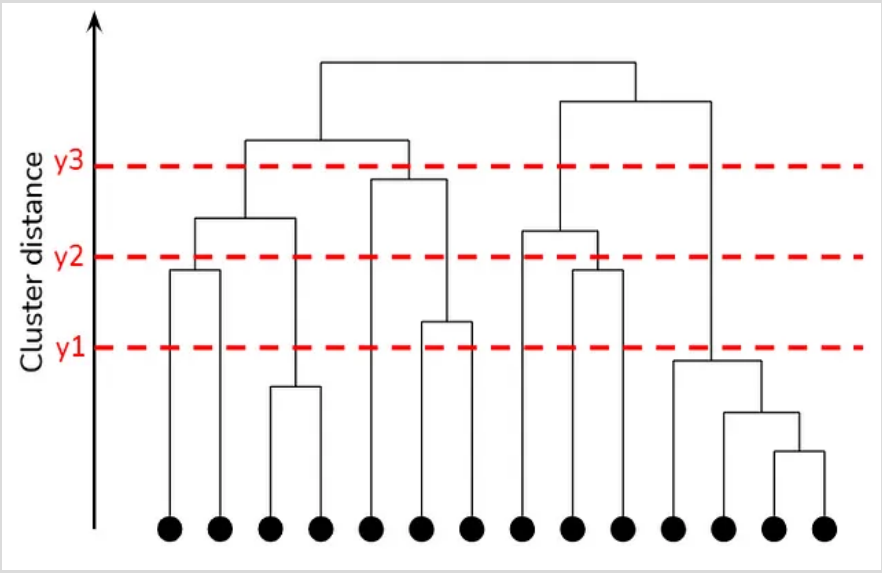
Process typically visualized using a tree-like diagram called dendrogram, which illustrates arrangement of clusters produced.

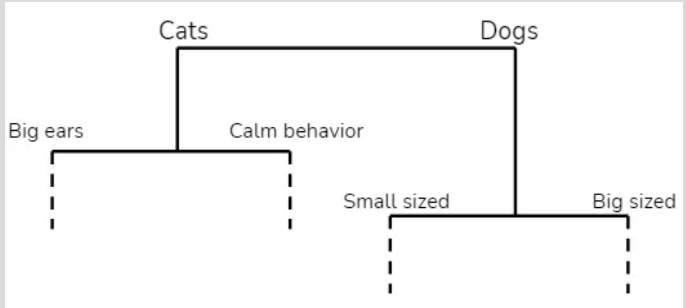
Determining Cluster Number

Number of clusters not predetermined; it's decided by cutting dendrogram at desired level, which gives flexible clustering solution.

Application Scenarios

- Ideal for data where structure unknown
- for exploratory data analysis, such as identifying species taxonomies or customer segments.





Density-Based Clustering

Groups together data points that are closely packed together, marking low-density regions where points are separated as the boundaries between clusters.

DBSCAN: A Popular Method

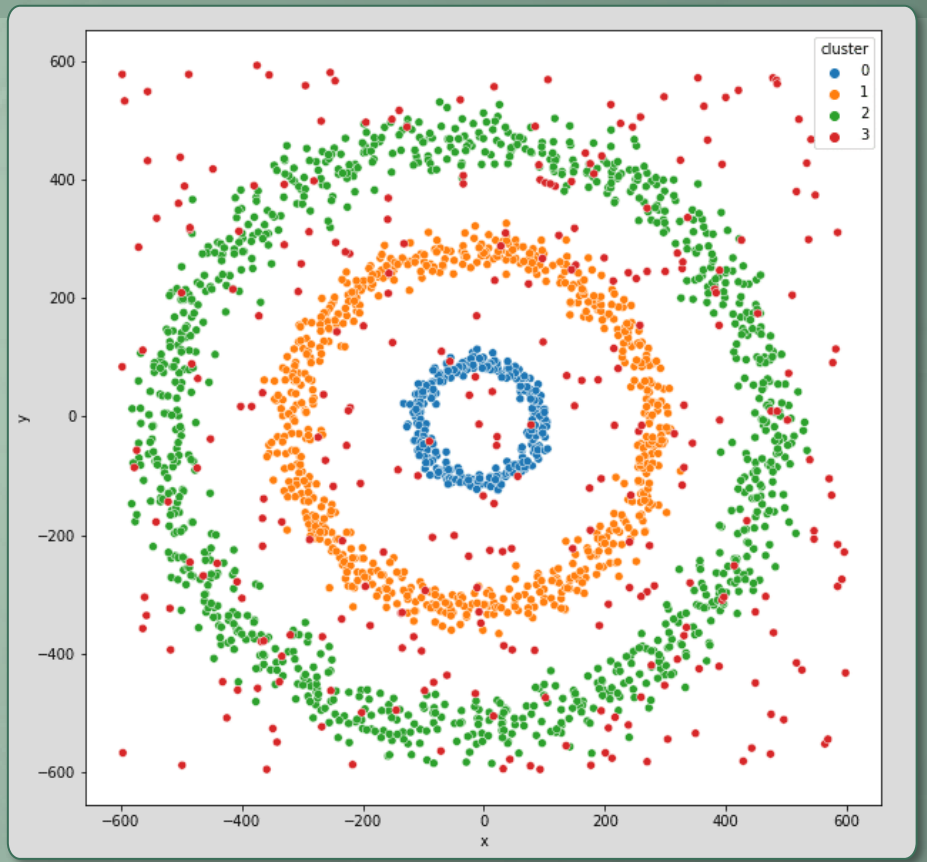
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters as high-density areas separated by areas of low density and is robust to noise.

Advantages of Density-Based Clustering

- Capable of finding arbitrarily shaped clusters.
- Good at separating noise from clusters.
- Does not require the number of clusters as input.

Usage Considerations

- Parameter selection crucial, it can significantly affect the outcome.
- Performs well in cases of spatial data clustering and can be applied to anomaly detection.



BIRCH Clustering

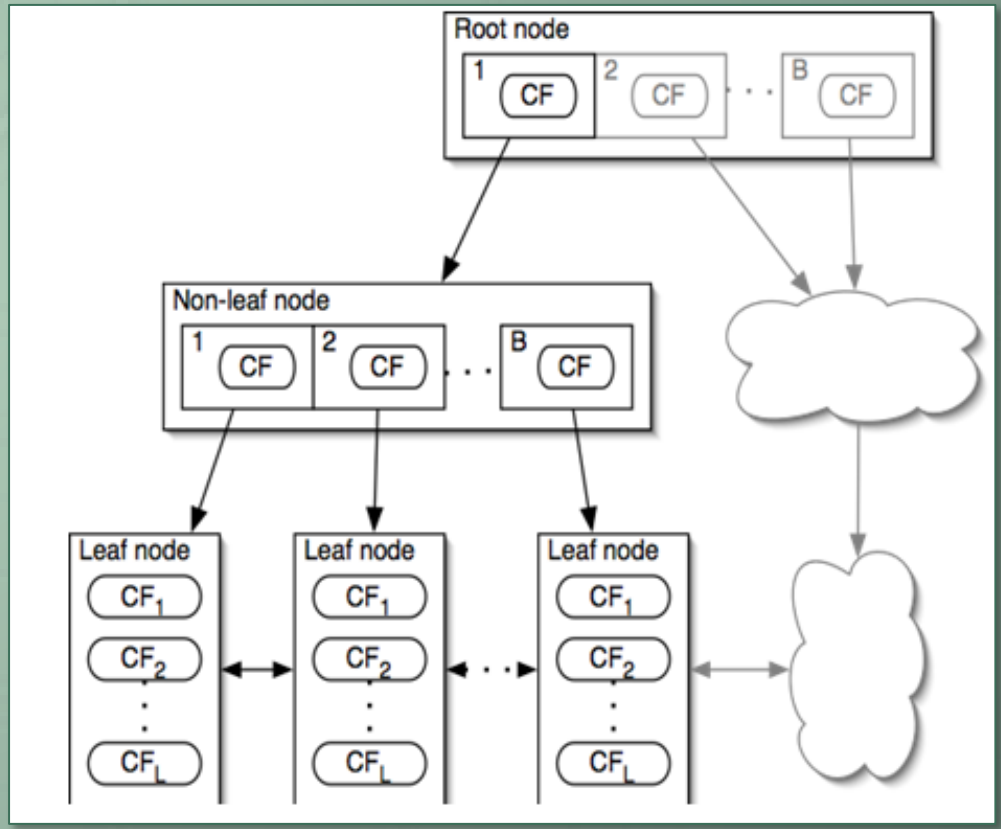
… is unsupervised data mining algorithm used for clustering large datasets

It means

- Balanced
- Iterative
- Reducing and
- Clustering using
- Hierarchies

Starting point: Clustering Feature Tree (CF-Tree)

- represents dataset with each leaf node entry representing subcluster
- Nonleaf nodes contain up to B entries, each with a pointer to a child node and a Clustering Feature (CF) summarizing the child's subclusters
- Leaf nodes hold up to L subclusters, with each one ensuring diameter of each subcluster is less than defined threshold



BIRCH Clustering

How the algorithm works

Phase 1: Initial Tree Building and Memory Management

- Starts with initial threshold, inserting points into CF-tree.
- If memory runs out, threshold is increased to rebuild smaller CF-tree.
- Re-inserts old leaf entries into new CF-tree, resuming data scanning from interruption point.

Phase 2 (optional): Subcluster Grouping

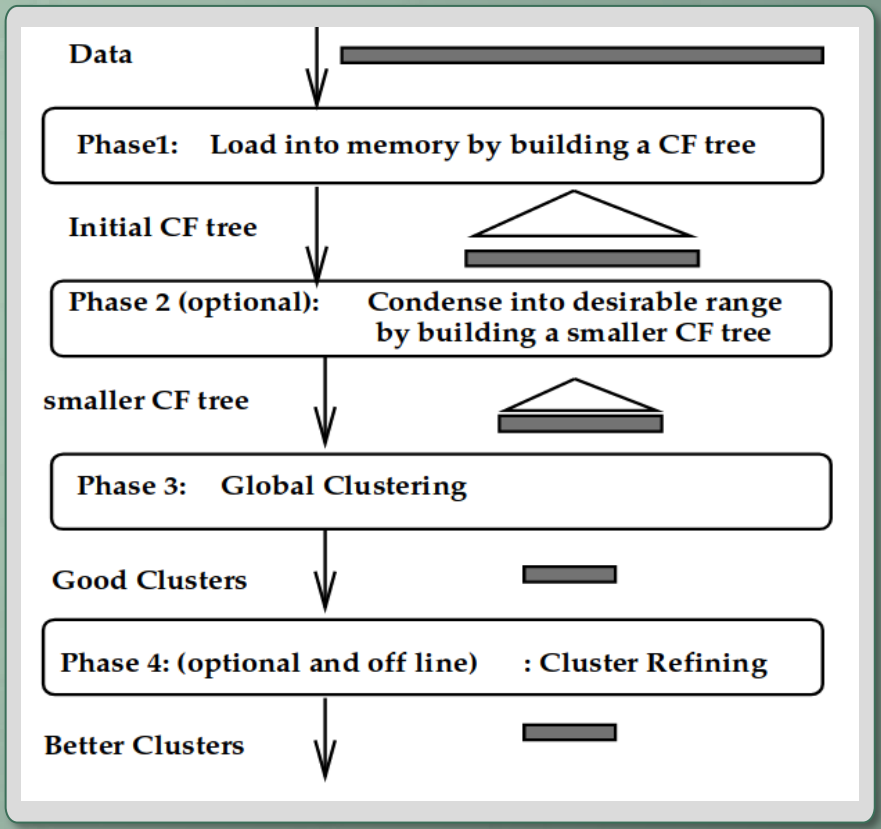
Groups crowded subclusters into larger ones for a more compact CF-tree.

Phase 3: Application of Secondary Clustering Algorithm

Adapts other algorithms (like k-MEANS) to categorize Clustering Features, maintaining BIRCH's efficiency.

Phase 4 (optional): Refinement and Outlier Discarding

Option to discard outliers and refine clustering based on detailed data analysis.



Comparison of clustering algorithms

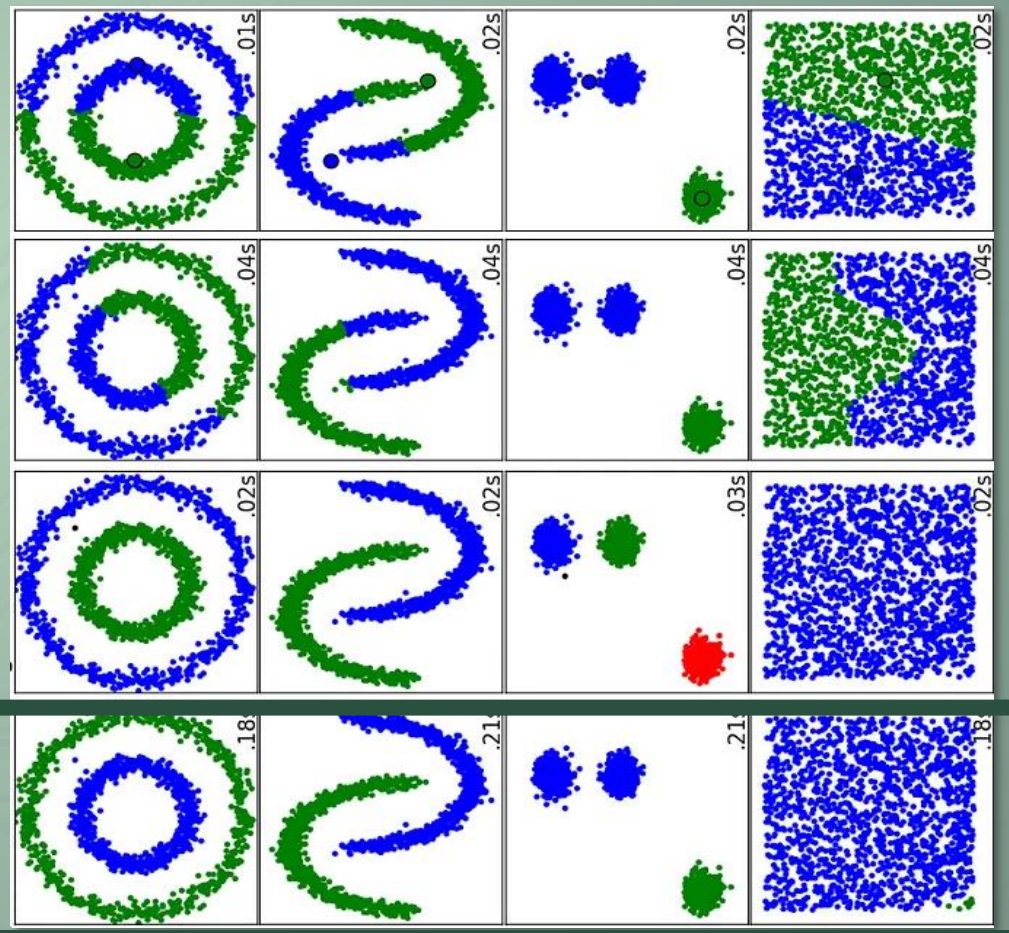
Centroid, hierarchy or density based algorithms perform well or not so well depending on structure and size of data

k-MEANS
- Centroid based
- Simple, however performance, reproducibility middle-rate

Hierarchical clustering
- Performance depends on data structure, good reproducibility
- Rather for small datasets

DBSCAN
- Density based
- Good performance also for complicated structures, rather for small datasets

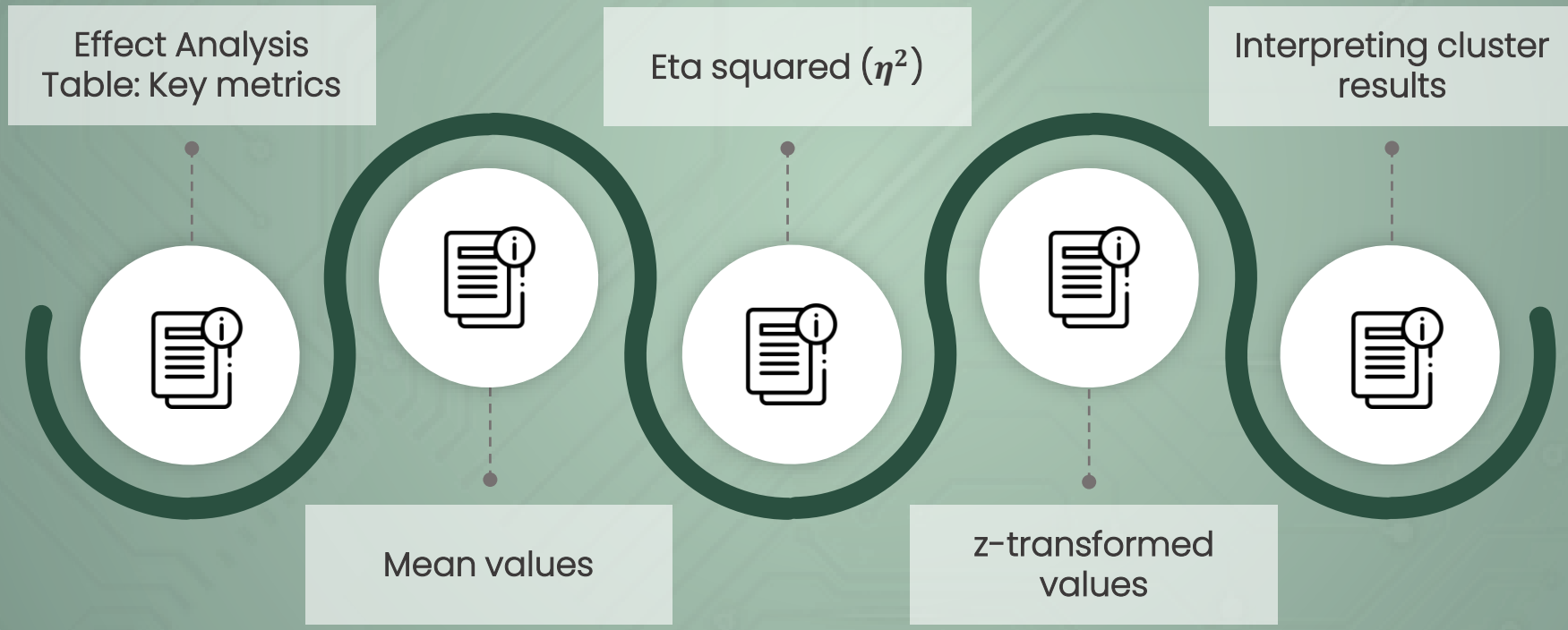
BIRCH
- Combination of hierarchy and centroid (k-means) based
- Good performance, reproducibility also for larger datasets



Focus

Effect Analysis Tables

… are used to quantify and display magnitude, significance of effects in experimental data and to describe and compare calculated clusters



Effect Analysis Table: Key metrics

Interpreting Mean Values, Eta Squared and Z-Transformations for all segmentation features

Mean Values

- Central measure indicating average outcomes within each group
- Fundamental for identifying scales and differences

Eta Squared (η^2)

- Measures proportion of total variance in dependent variable that is attributable to a factor,
- Key for understanding effect size in ANOVA tests

Z-Transformed Values

- Standardization of data points based on mean, standard deviation
- Useful for comparing scores from different distributions and identifying outliers

Index	description	1	2	3	4	otal mea	Eta squarec	z_1	z_2	z_3	z_4
0	Population ratio: school diploma	23.5387	2.54316	11.5277	5.43047	4.89768	0.886345	3.8453	-0.485695	1.36765	0.109903
1	Population ratio: college degree	35.4876	4.19433	16.1253	7.77358	7.38703	0.874354	4.01073	-0.455686	1.24719	0.0551722
2	Population in the house	74.7547	10.8867	37.1235	18.1483	17.6136	0.851693	3.85619	-0.453966	1.31664	0.0360881
3	Population ratio: university degree	17.2787	2.46361	9.34252	4.64443	4.21684	0.774792	3.46458	-0.465035	1.35956	0.113415
4	Tendency of having children	7.3713	3.36733	7.5142	6.21051	4.564	0.730642	1.44134	-0.6144	1.51471	0.845363
5	Interest: social media	6.84566	4.0987	6.96723	6.85288	5.08522	0.566425	0.990565	-0.555095	1.05897	0.994629
6	House size: number of flats	2.37237	1.31871	3.11911	2.56002	1.81605	0.564792	0.601959	-0.538126	1.40994	0.804996
7	Interest: green energy	8.29343	5.7935	8.80948	8.966	6.87341	0.506247	0.688072	-0.523269	0.938124	1.01397
8	Main purchase criterion: low price	7.78607	5.10585	7.99666	8.19596	6.16072	0.439162	0.752138	-0.488142	0.84959	0.941815
9	Population density (households per qkm)	8.00399	5.56313	10.599	9.90372	7.11272	0.438949	0.275552	-0.479079	1.07785	0.862881
10	Interest: online shopping	5.44948	4.56909	6.50069	6.69782	5.2555	0.300827	0.111423	-0.39428	0.715252	0.828487
11	Status of Location	2.80903	2.05673	3.00239	2.73242	2.32961	0.177814	0.536042	-0.305113	0.752239	0.450386
12	Interest: politics	6.8313	6.29871	7.65781	8.1999	6.8604	0.174146	-0.0151308	-0.292022	0.414574	0.696403
13	Interest: football	3.33846	2.91265	4.27357	3.76197	3.25778	0.168634	0.0662808	-0.283511	0.834444	0.414179
14	Interest: travel	6.13116	6.33569	6.71202	7.83844	6.67687	0.085086	-0.265093	-0.165737	0.0170743	0.564263
15	Tendency of having a photovoltaic system	3.3949	3.48795	3.35546	4.39615	3.65331	0.0582588	-0.165109	-0.105654	-0.190305	0.474621
16	Interest: economics	5.91435	5.16395	5.28867	5.81669	5.33658	0.0507714	0.465909	-0.139203	-0.038631	0.387153
17	Main purchase criterion: brand/quality	4.32879	3.90469	3.96371	4.4824	4.04322	0.0280004	0.20326	-0.0986019	-0.0565912	0.312599
18	Interest: alternative energy	3.77087	3.87738	3.87853	4.48164	3.99626	0.0240819	-0.142215	-0.0750053	-0.0742799	0.306257

Effect Analysis: Mean values

… serve as a central reference point to summarize data within each cluster, indicating average characteristics of clustered groups.

Role of Mean Values in Clustering

- Cluster Profiling:
  Mean values help define the profile of each cluster, enabling clear understanding of group characteristics.
- Scale Identification:
  Scale differences of features become visible.
- Model drift monitoring:
  Track changes in mean values over time, i.e. model drift to monitor cluster performance and stability.

Utilizing Mean Values for Decision-Making

- Targeted Strategy Development: Inform the development of strategies tailored to the specific traits of each cluster.

Focus

Index	description	1	2	3	4	otal mea
0	Population ratio: school diploma	23.5387	2.54316	11.5277	5.43047	4.89768
1	Population ratio: college degree	35.4876	4.19433	16.1253	7.77358	7.38703
2	Population in the house	74.7547	10.8867	37.1235	18.1483	17.6136
3	Population ratio: university degree	17.2787	2.46361	9.34252	4.64443	4.21684
4	Tendency of having children	7.3713	3.36733	7.5142	6.21051	4.564
5	Interest: social media	6.84566	4.0987	6.96723	6.85288	5.08522
6	House size: number of flats	2.37237	1.31871	3.11911	2.56002	1.81605
7	Interest: green energy	8.29343	5.7935	8.80948	8.966	6.87341
8	Main purchase criterion: low price	7.78607	5.10585	7.99666	8.19596	6.16072
9	Population density (households per qkm)	8.00399	5.56313	10.599	9.90372	7.11272
10	Interest: online shopping	5.44948	4.56909	6.50069	6.69782	5.2555
11	Status of Location	2.80903	2.05673	3.00239	2.73242	2.32961
12	Interest: politics	6.8313	6.29871	7.65781	8.1999	6.8604
13	Interest: football	3.33846	2.91265	4.27357	3.76197	3.25778
14	Interest: travel	6.13116	6.33569	6.71202	7.83844	6.67687
15	Tendency of having a photovoltaic system	3.3949	3.48795	3.35546	4.39615	3.65331
16	Interest: economics	5.91435	5.16395	5.28867	5.81669	5.33658
17	Main purchase criterion: brand/quality	4.32879	3.90469	3.96371	4.4824	4.04322
18	Interest: alternative energy	3.77087	3.87738	3.87853	4.48164	3.99626

Effect Analysis: Eta Squared (η^2)

… measures effect size of input features on separability of clusters.
The most important input feature for separability of clusters has largest η^2

Significance of η^2 in Analysis

- Variance Explanation: Reflects strength of association between variables, with higher η^2 values denoting stronger relationship.
- Here this relationship refers to separability of clusters

Calculating Eta Squared

- Calculated as the sum of squares between groups (SSB) over the total sum of squares (SST): η^2 = SSB/SST.

Interpretation: Clusters are majorly distinguished by

- Population ratio of different educational degrees
- Number of population in houses
- Tendency of having children, i.e. family sizes
- Interest or no interest in social media

Index	description	Eta squared
0	Population ratio: school diploma	0.867258
1	Population ratio: college degree	0.853004
2	Population in the house	0.764085
3	Population ratio: university degree	0.72884
4	Tendency of having children	0.300089
5	Interest: social media	0.240682
6	House size: number of flats	0.222522
7	Interest: green energy	0.138611
8	Main purchase criterion: low price	0.115523
9	Population density (households per qkm)	0.0764655
10	Interest: online shopping	0.0551705
11	Status of Location	0.0540565
12	Interest: politics	0.0450985
13	Interest: football	0.0363586
14	Interest: travel	0.0274985
15	Tendency of having a photovoltaic syste	0.0273676
16	Interest: economics	0.0204846
17	Main purchase criterion: brand/quality	0.0165858
18	Interest: alternative energy	0.0102784

Focus

Effect Analysis: z-transformed Values

… measure standard deviation that a data point is from mean of its dataset which can be used as fingerprint of each cluster compared with other clusters

Fingerprint of cluster

Characteristics of each cluster compared with other clusters become visible

Purpose of z-transformed values in technical sense

- Normalization: Converts different data sets to a common scale without distorting differences in the ranges of values.
- Outlier Identification: Facilitates the recognition of outliers by highlighting values significantly distant from the mean.

Computing Z-Transformed Values

Calculated using the formula: $Z = \frac{(X-\mu)}{\sigma}$, where X is a data point, μ is the mean, and σ is the standard deviation.

Index	description	z_1	z_2	z_3	z_4
0	Population ratio: school diploma	3.8453	-0.485695	1.36765	0.109903
1	Population ratio: college degree	4.01073	-0.455686	1.24719	0.0551722
2	Population in the house	3.85619	-0.453966	1.31664	0.0360881
3	Population ratio: university degree	3.46458	-0.465035	1.35956	0.113415
4	Tendency of having children	1.44134	-0.6144	1.51471	0.845363
5	Interest: social media	0.990565	-0.555095	1.05897	0.994629
6	House size: number of flats	0.601959	-0.538126	1.40994	0.804996
7	Interest: green energy	0.688072	-0.523269	0.938124	1.01397
8	Main purchase criterion: low price	0.752138	-0.488142	0.84959	0.941815
9	Population density (households per qkm)	0.275552	-0.479079	1.07785	0.862881
10	Interest: online shopping	0.111423	-0.39428	0.715252	0.828487
11	Status of Location	0.536042	-0.305113	0.752239	0.450386
12	Interest: politics	-0.0151308	-0.292022	0.414574	0.696403
13	Interest: football	0.0662808	-0.283511	0.834444	0.414179
14	Interest: travel	-0.265093	-0.165737	0.0170743	0.564263
15	Tendency of having a photovoltaic system	-0.165109	-0.105654	-0.190305	0.474621
16	Interest: economics	0.465909	-0.139203	-0.038631	0.387153
17	Main purchase criterion: brand/quality	0.20326	-0.0986019	-0.0565912	0.312599
18	Interest: alternative energy	-0.142215	-0.0750053	-0.0742799	0.306257

Focus

Interpreting cluster descriptions

… using most extreme, i.e. most positive or negative z-transformed values is useful tool for targeted marketing among existing customers

Cluster 1:
Educated urban families

- People with higher education degree
- Families with kids
- Living in larger buildings, urban
- Being interested in social media

Cluster 2:
Brand aware working class

- Working class
- Interested in photovoltaics, travelling
- Brand and quality aware

Focus

Index	description	z_1	z_2	z_3	z_4
0	Population ratio: school diploma	3.8453	-0.485695	1.36765	0.109903
1	Population ratio: college degree	4.01073	-0.455686	1.24719	0.0551722
2	Population in the house	3.85619	-0.453966	1.31664	0.0360881
3	Population ratio: university degree	3.46458	-0.465035	1.35956	0.113415
4	Tendency of having children	1.44134	-0.6144	1.51471	0.845363
5	Interest: social media	0.990565	-0.555095	1.05897	0.994629
6	House size: number of flats	0.601959	-0.538126	1.40994	0.804996
7	Interest: green energy	0.688072	-0.523269	0.938124	1.01397
8	Main purchase criterion: low price	0.752138	-0.488142	0.84959	0.941815
9	Population density (households per qkm)	0.275552	-0.479079	1.07785	0.862881
10	Interest: online shopping	0.111423	-0.39428	0.715252	0.828487
11	Status of Location	0.536042	-0.305113	0.752239	0.450386
12	Interest: politics	-0.0151308	-0.292022	0.414574	0.696403
13	Interest: football	0.0662808	-0.283511	0.834444	0.414179
14	Interest: travel	-0.265093	-0.165737	0.0170743	0.564263
15	Tendency of having a photovoltaic system	-0.165109	-0.105654	-0.190305	0.474621
16	Interest: economics	0.465909	-0.139203	-0.038631	0.387153
17	Main purchase criterion: brand/quality	0.20326	-0.0986019	-0.0565912	0.312599
18	Interest: alternative energy	-0.142215	-0.0750053	-0.0742799	0.306257

Interpreting cluster descriptions

… using most extreme, i.e. most positive or negative z-transformed values is useful tool for targeted marketing among existing customers

Cluster 3:
Upper-class Families interested in sports

- People with college degree
- Families with kids
- Living in small houses of higher status
- Interested in sports

Cluster 4:
Low-income suburbian families

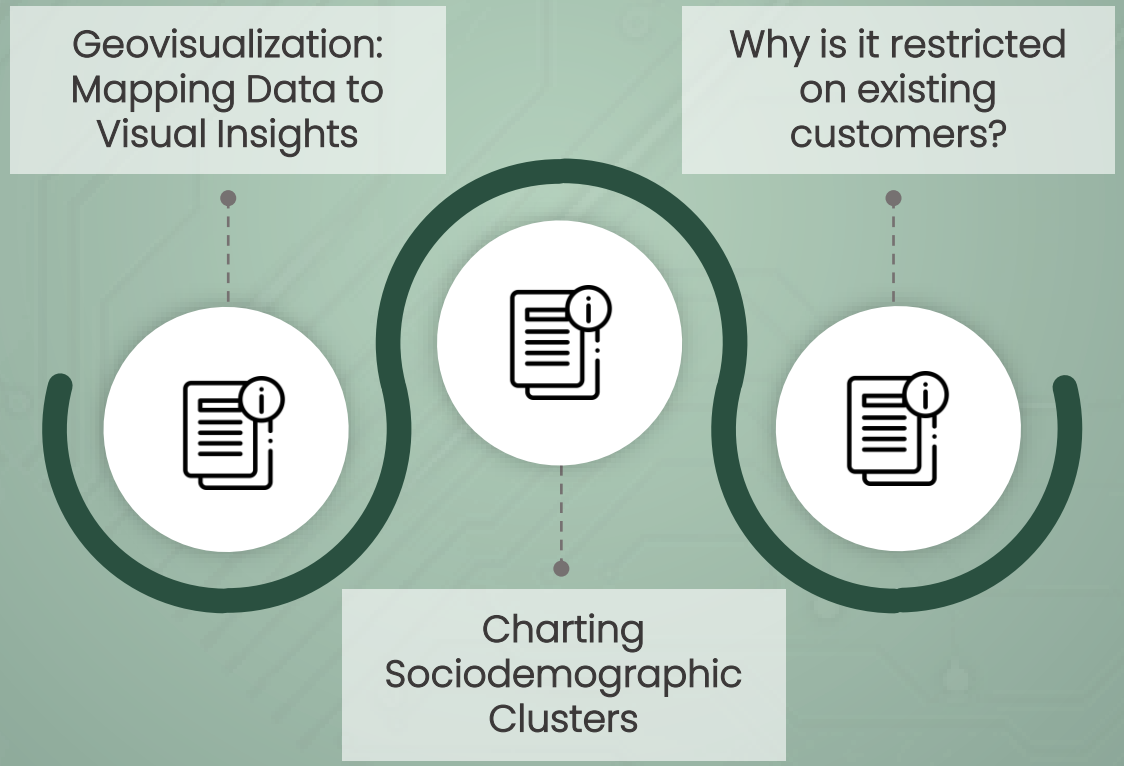
- Low population density, i.e. suburbian
- Families with kids
- Low price oriented, rather poor

Focus

Index	description	z_1	z_2	z_3	z_4
0	Population ratio: school diploma	3.8453	-0.485695	1.36765	0.109903
1	Population ratio: college degree	4.01073	-0.455686	1.24719	0.0551722
2	Population in the house	3.85619	-0.453966	1.31664	0.0360881
3	Population ratio: university degree	3.46458	-0.465035	1.35956	0.113415
4	Tendency of having children	1.44134	-0.6144	1.51471	0.845363
5	Interest: social media	0.990565	-0.555095	1.05897	0.994629
6	House size: number of flats	0.601959	-0.538126	1.40994	0.804996
7	Interest: green energy	0.688072	-0.523269	0.938124	1.01397
8	Main purchase criterion: low price	0.752138	-0.488142	0.84959	0.941815
9	Population density (households per qkm)	0.275552	-0.479079	1.07785	0.862881
10	Interest: online shopping	0.111423	-0.39428	0.715252	0.828487
11	Status of Location	0.536042	-0.305113	0.752239	0.450386
12	Interest: politics	-0.0151308	-0.292022	0.414574	0.696403
13	Interest: football	0.0662808	-0.283511	0.834444	0.414179
14	Interest: travel	-0.265093	-0.165737	0.0170743	0.564263
15	Tendency of having a photovoltaic system	-0.165109	-0.105654	-0.190305	0.474621
16	Interest: economics	0.465909	-0.139203	-0.038631	0.387153
17	Main purchase criterion: brand/quality	0.20326	-0.0986019	-0.0565912	0.312599
18	Interest: alternative energy	-0.142215	-0.0750053	-0.0742799	0.306257

Geovisualization

... is about visualizing geographic data through maps and other graphical representations to analyze and communicate spatial information effectively



Geovisualization: Mapping Data to Visual Insights

Geovisualization integrates data analysis and interactive mapping to transform spatial data into visual representations.

Elements of Geovisualization

- Spatial Representation:
  Use of maps as base to overlay data points, lines, and polygons.
- Thematic Layering:
  Applying layers of data, such as population density or climate patterns, to provide context and enhance understanding.

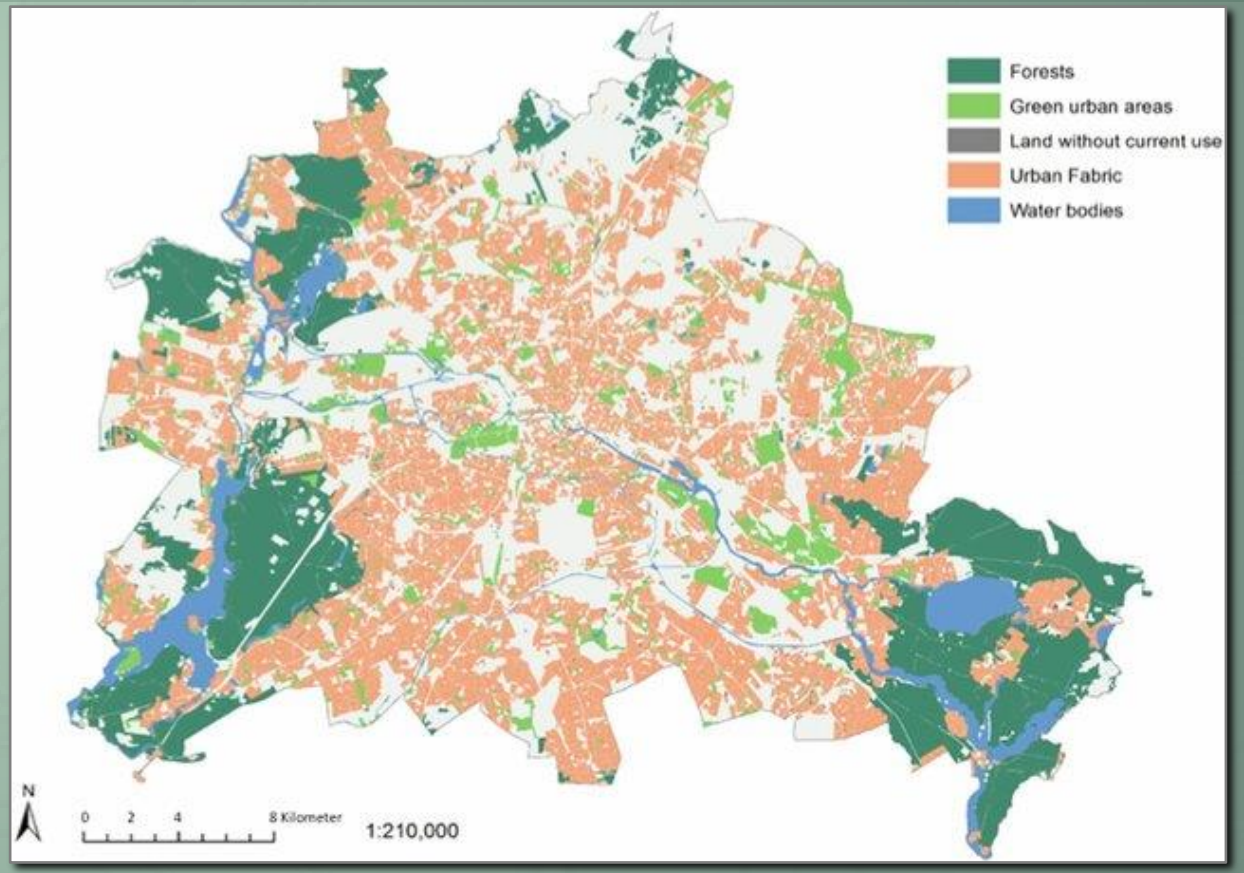
Benefits of Geovisualization

- Intuitive Analysis:
  Offers an immediate, visual form of data interpretation that can be more accessible than raw numbers.
- Pattern Recognition:
  Aids in identifying trends, anomalies, and correlations within geographically-tied data.

Applications Across Fields, used in ...

- Urban planning
- Environmental monitoring
- Market analysis and more

... to support decision-making processes.



Charting Sociodemographic Clusters

… is about depicting spatial distribution of population segments defined by distinguishing characteristics, revealing patterns, i.e. as basis for marketing strategies

Key Components in Sociodemographic Mapping

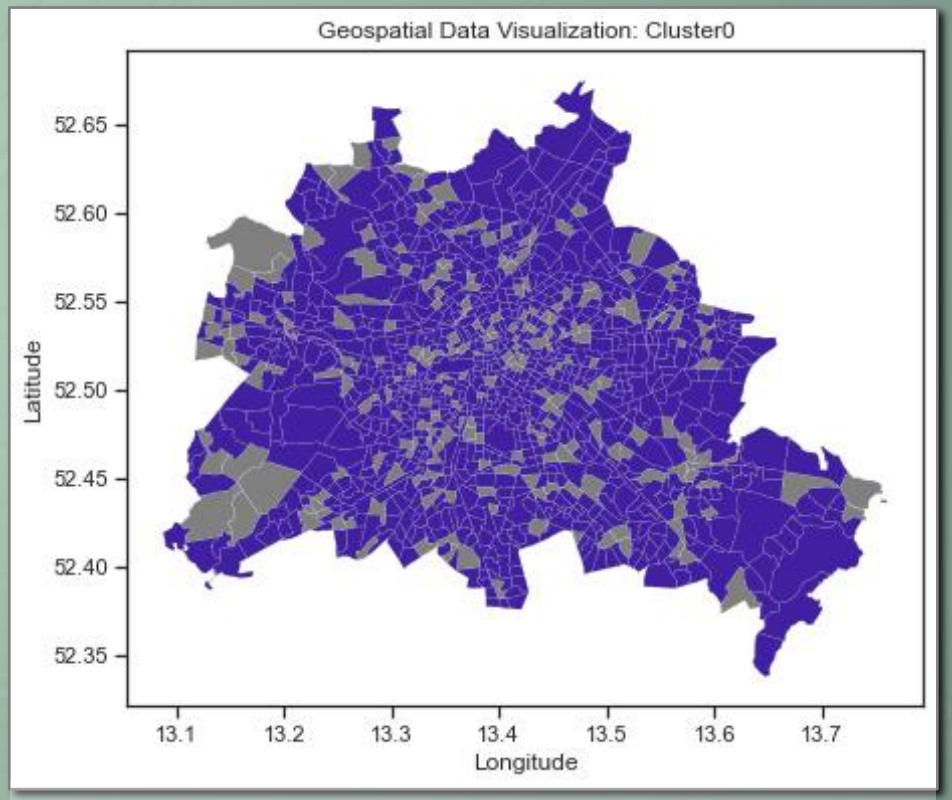
- Data Integration: Combines census data, consumer behavior, and geographical information.
- Cluster Visualization: Utilizes color coding and symbols to differentiate clusters based on sociodemographic traits.

Insights Gained from Sociodemographic Maps

- Community Profiling: Identifies community characteristics for tailored public services and policy-making.
- Market Segmentation: Enhances targeted marketing efforts by locating and understanding consumer segments.

Challenges and Considerations

- Balancing detail and readability, ensuring privacy, and addressing the dynamic nature of sociodemographic data.



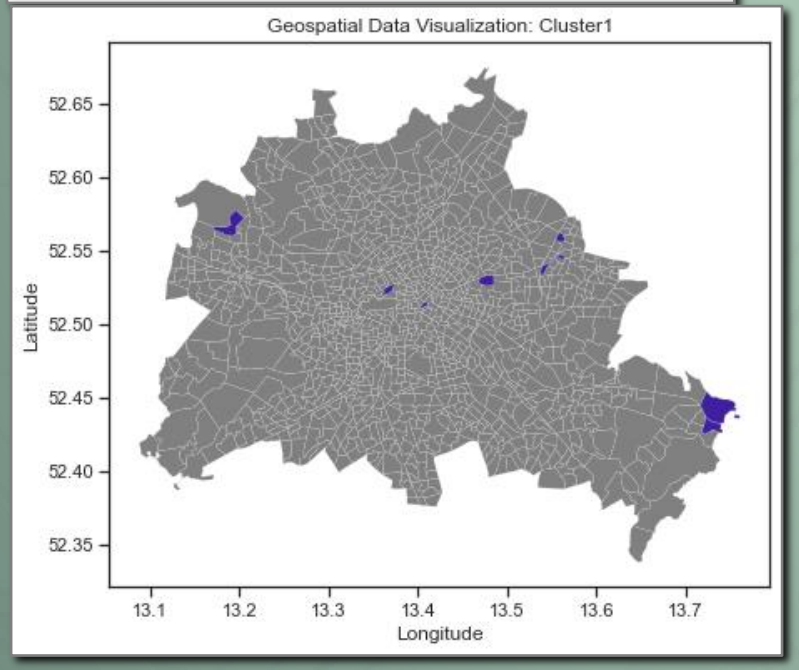
Marked: Geolocations without sales events

Charting Sociodemographic Clusters

... in order to depict the spatial distribution of population characteristics, revealing patterns that may influence marketing strategies.

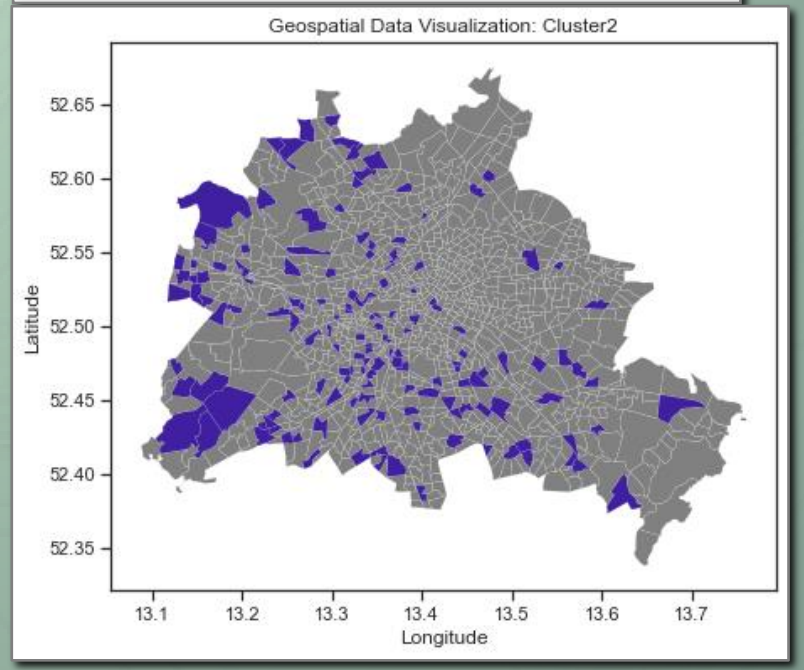
Cluster 1:
Educated urban families

- People with higher education degree
- Families with kids
- Living in larger buildings, urban
- Being interested in social media



Cluster 2:
Brand aware working class

- Working class
- Interested in photovoltaics, travelling
- Brand and quality aware

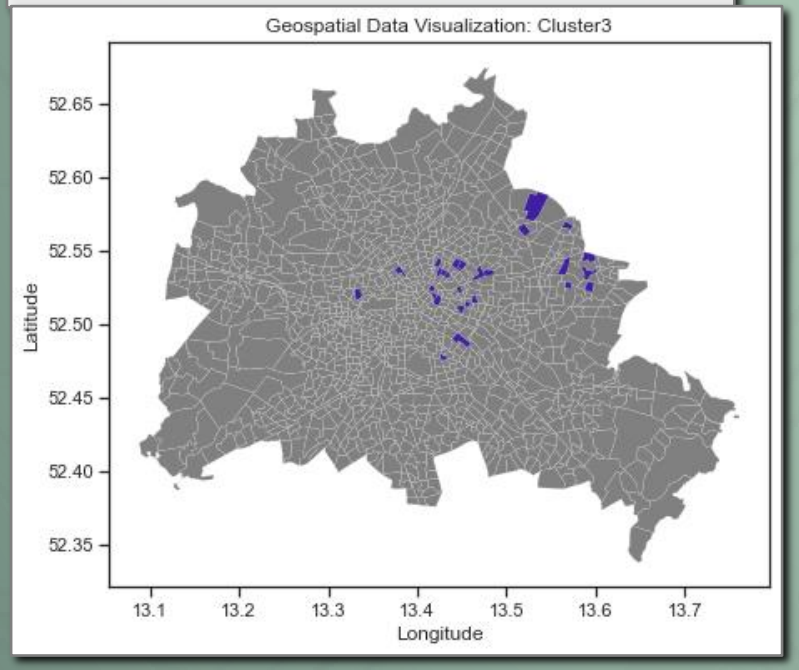


Charting Sociodemographic Clusters

… in order to depict the spatial distribution of population characteristics, revealing patterns that may influence marketing strategies.

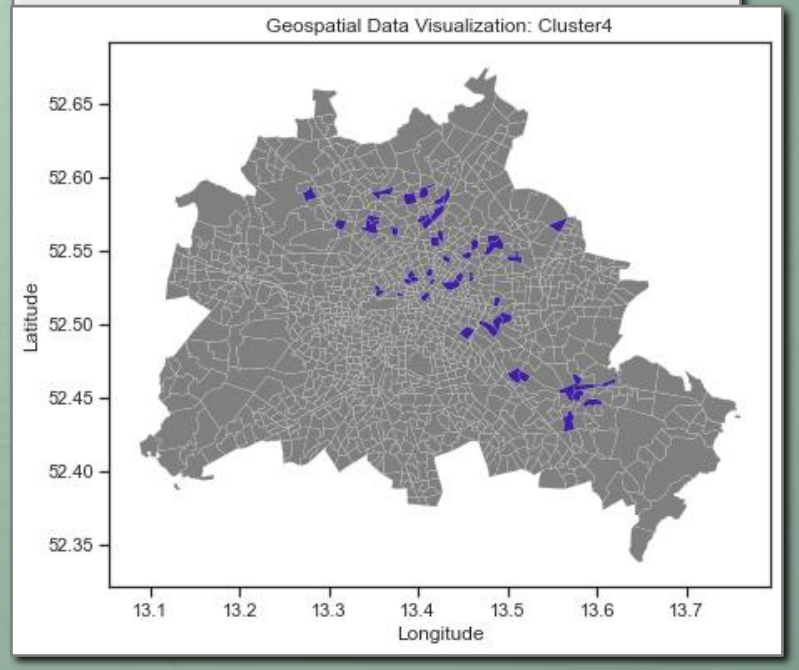
Cluster 3:
Upper-class Families interested in sports

- People with college degree
- Families with kids
- Living in small houses of higher status
- Interested in sports



Cluster 4:
Low-income suburbian families

- Low population density, i.e. suburbian
- Families with kids
- Low price oriented, rather poor



Links

i.e. sources for self-learning

	Title	Link
Factor analysis	Principal Component Analysis(PCA)	https://www.geeksforgeeks.org/principal-component-analysis-pca/
	PCA : how to interpret the weights/loadings and Varimax rotation	https://www.youtube.com/watch?v=BiuwDI_BbWw

	Title	Link
Clustering	Unveiling the Patterns: A Journey into Clustering Algorithms	https://medium.com/@sharaffinb/unveiling-the-patterns-a-journey-into-clustering-algorithms-3ed7d40eb1de
	Understanding Cluster Analysis: Data Patterns and Relationships	https://medium.com/@mondoa/understanding-cluster-analysis-data-patterns-and-relationships-4072c6715614
	Hierarchical clustering explained	https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8
	10 Clustering Algorithms With Python	https://machinelearningmastery.com/clustering-algorithms-with-python/
	Different Types of Clustering Methods in Unsupervised Learning	https://pub.towardsai.net/machine-learning-16c8ccc2c7b8

	Title	Link
Effect sizes	Python - Eta squared	https://www.youtube.com/watch?v=eycsylkMJNw
	13. Effect size in ANOVA Eta squared	https://www.youtube.com/watch?app=desktop&v=NQDIm7pn3Pk

Links

i.e. sources for self-learning

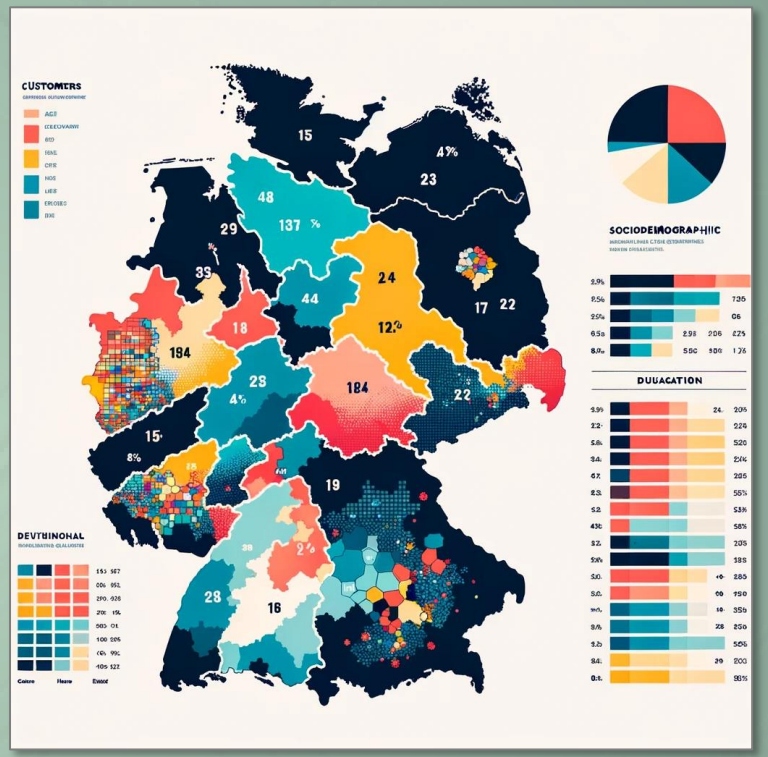
	Title	Link
Clustering: DBCAN	DBSCAN in Python: learn how it works	https://anderfernandez.com/en/blog/dbscan-python-tutorial/
	How DBSCAN works and why should we use it?	https://medium.com/@mondoa/understanding-cluster-analysis-data-patterns-and-relationships-4072c6715614
	How to Master the Popular DBSCAN Clustering Algorithm for Machine Learning	https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

	Title	Link
Clustering: BIRCH	BIRCH Clustering Algorithm Example In Python	https://towardsdatascience.com/machine-learning-birch-clustering-algorithm-clearly-explained-fb9838cbeed9
	Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) in Artificial intelligence	https://medium.com/@bharathwajan/balanced-iterative-reducing-and-clustering-using-hierarchies-birch-in-artificial-intelligence-6a3f125c657f
	BIRCH Clustering Algorithm Example In Python	https://towardsdatascience.com/machine-learning-birch-clustering-algorithm-clearly-explained-fb9838cbeed9
	BALANCED ITERATIVE REDUCING AND CLUSTERING USING HEIRARCHIES(BIRCH)	https://medium.com/@noel.cs21/balanced-iterative-reducing-and-clustering-using-heirachies-birch-5680adffaa58

	Title	Link
Geo-visualization	In search of features that constitute an “enriched environment” in humans: Associations between geographical properties and brain structure	https://www.researchgate.net/publication/319940680_In_search_of_features_that_constitute_an_enriched_environment_in_humans_Associations_between_geographical_properties_and_brain_structure

ChatGPT/Dall-E3 Prompts

Create a photo of a data scientist, a woman with shoulder-length curly hair, focused on analyzing a large, detailed heat map of Germany displayed on a wall-mounted screen. The heat map transitions from a green to gray gradient to indicate varying degrees of sales intensity across different regions, with clear annotations for key metrics. She is taking notes on a digital tablet, referencing the heat map for sales prediction.



Design a simple image with a map of Germany as the base, segmenting customers sociodemographically. Overlay the map with colored regions indicating different sociodemographic groups such as age, income, and education level. Use distinct colors for each group, and include a clear legend on the side of the map to detail what each color represents. The regions should be clearly outlined, and the text must be legible, allowing for easy interpretation of the data.

About me

Dr. Harald Stein

- Data Scientist ~ 7 years experience
- Algotrader ~ 4 years experience
- Ph.D. in Economics, Game Theory

- LinkedIn: https://www.linkedin.com/in/harald-stein-phd-1648b51a
- ResearchGate: https://www.researchgate.net/profile/Harald-Stein