

Datafest Munich 2025

Data tutorial and questions

March 28th, 2025



Contents

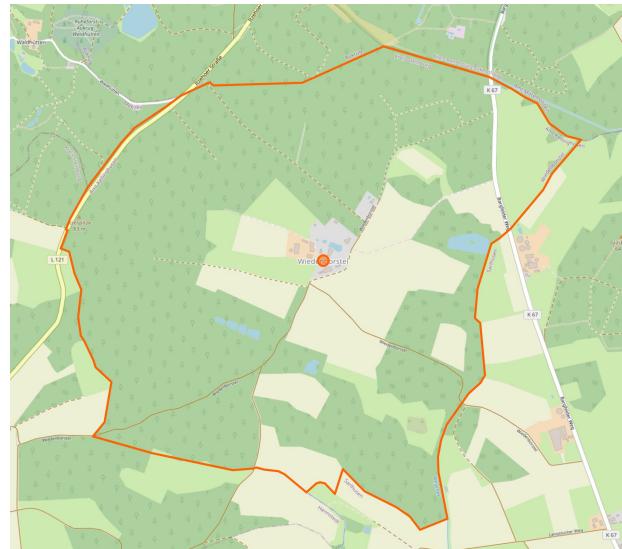
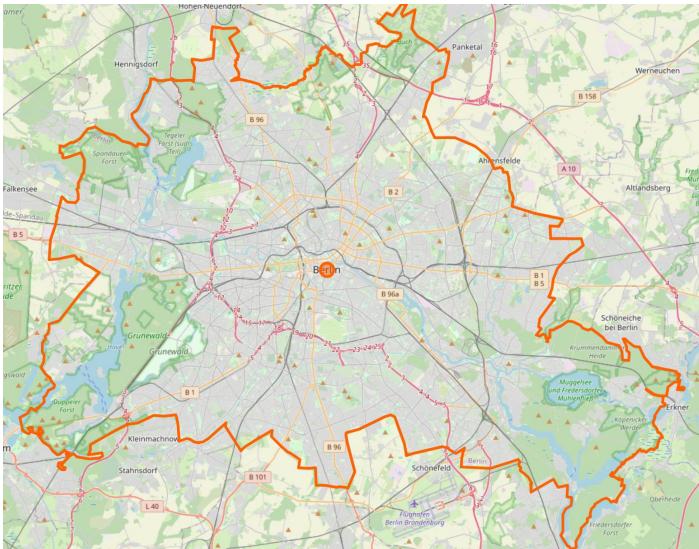
- Background of the data
- Working with geospatial data
- Linking to other datasets and data sources
- Coming up with questions: Ideation methods
- Example questions and approaches
- Questions

Background of the data

- Caveats/things to watch out:
 - Different data sources mean that temporal and spatial availability varies.
Some variables have a high share of NAs
 - We checked the data, but you might still find inconsistencies or matching errors -> if so please let us know
- Data use
 - The data is not proprietary or private. You don't need to delete it after Datafest 2025
 - BUT: please cite appropriately (esp FDZ Ruhr data) if you use it to publish

Background of the data

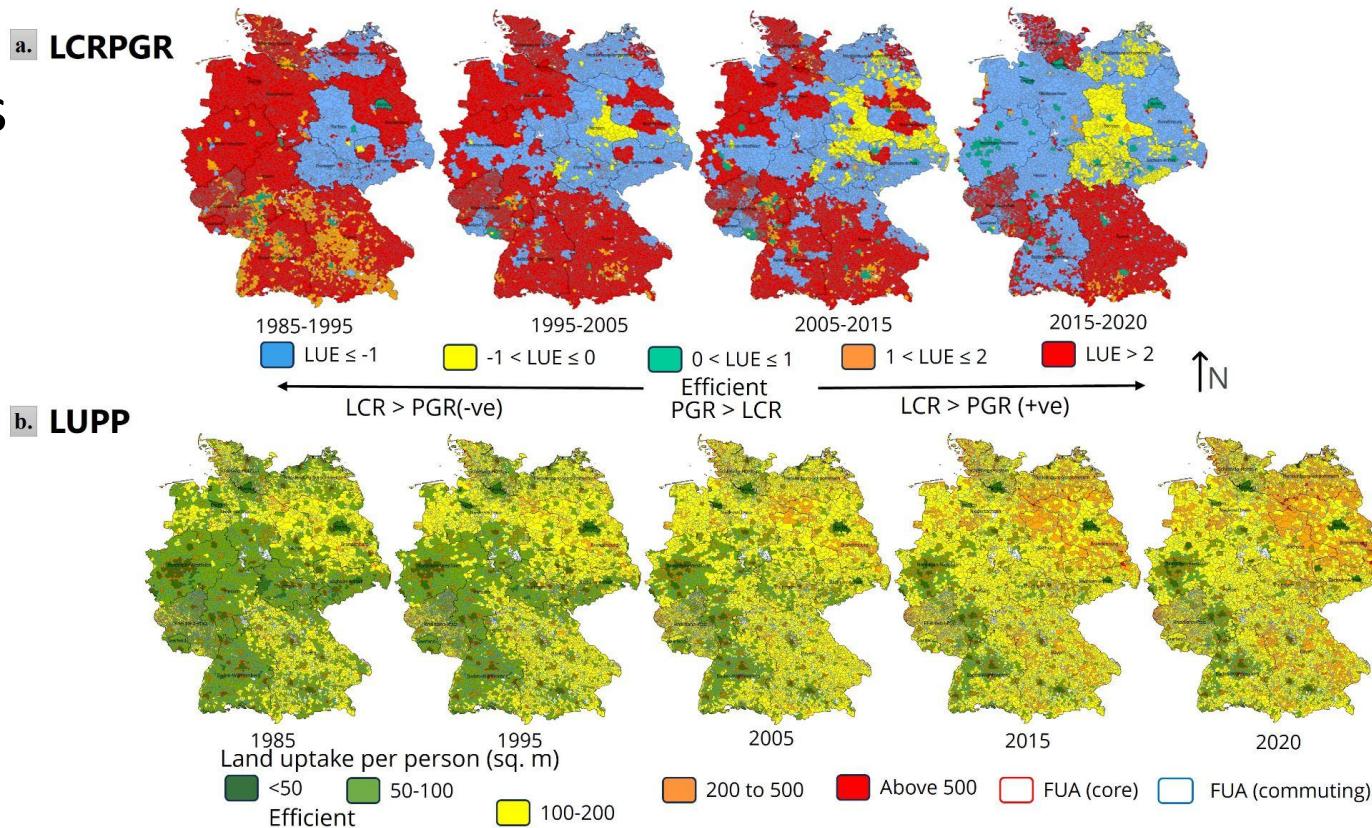
- Municipal data might need cleaning, pre-processing and selection before you begin the analysis
- All municipalities have the same administrative status, but there are large differences: Berlin (3.6 million people) and Wiedenborstel (12 people) are both municipalities



Sources: OSM

Background of the data: East/West and Urban/rural differences

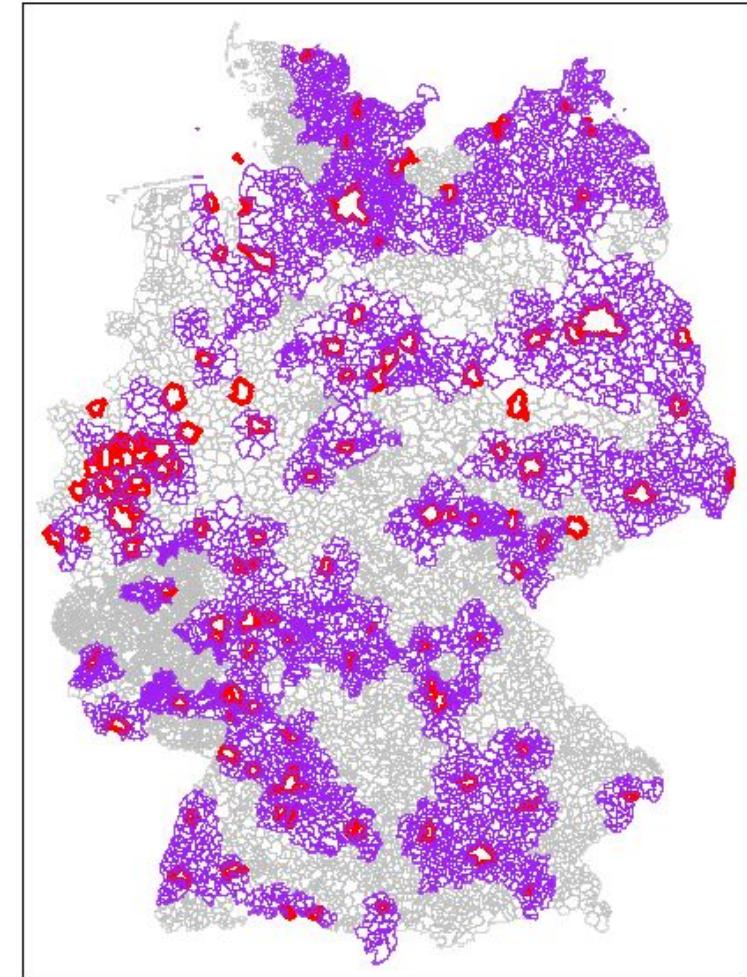
- Land use and population follow different trajectories in former East/West Germany and in urban/rural areas.
- To understand more subtle differences one might split the data based on those categories



Sources: Own calculations based on GHSL

Background of the data: Functional Urban Areas (FUA)

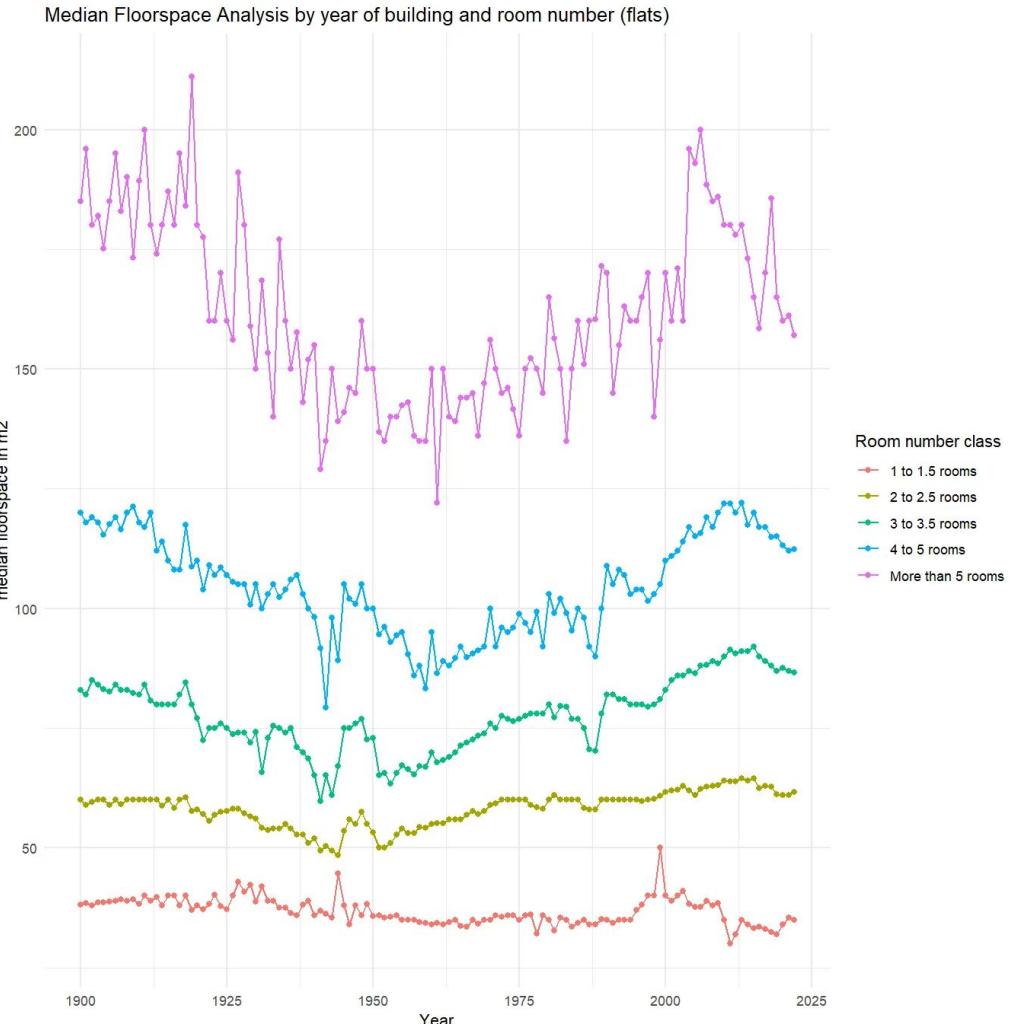
- FUA core zones are cities with high density built surfaces and a population of over 50000 (red)
- FUA commuting zones are adjacent municipalities, where at least 15% of the residents work in the core zone (purple)



Sources: Own mapping based on OECD, GHSL and BKG

Background: Housing Ads Data (FDZ)

- Data is based on scraped Immoscout24 ads and already pre-processed (e.g. yearly sampling)
 - Caveats: Immoscout ads might not be representative of all housing offers, prices are initial offers from the seller, not real closing contracts
- Disaggregated by housing type (e.g. detached house, block of flats, ...)
- Most consistent information on sale prices, rents, number of rooms, living space (Wohnfläche), building year...
- Other variables like heating type have lot of missing values
- Spatial matching possible by municipality code (AGS), grid cell (1km resolution) or postal code

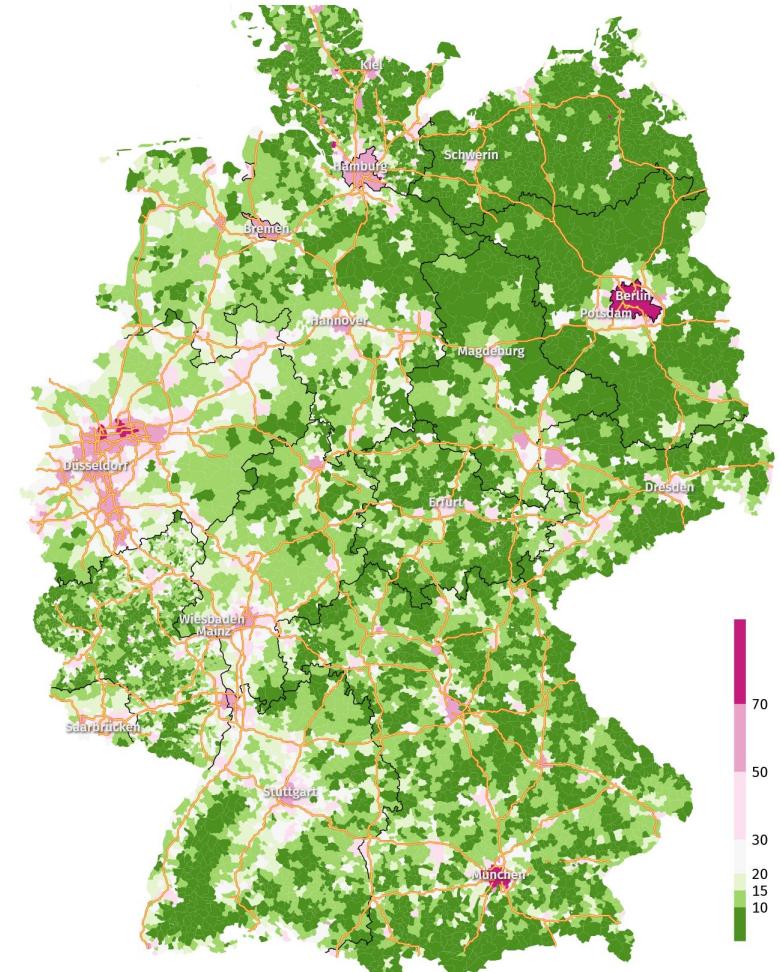


Sources: own material based on RWI-GEO-RED 2024

Working with geospatial data

Geospatial data is data that is linked with real-world locations

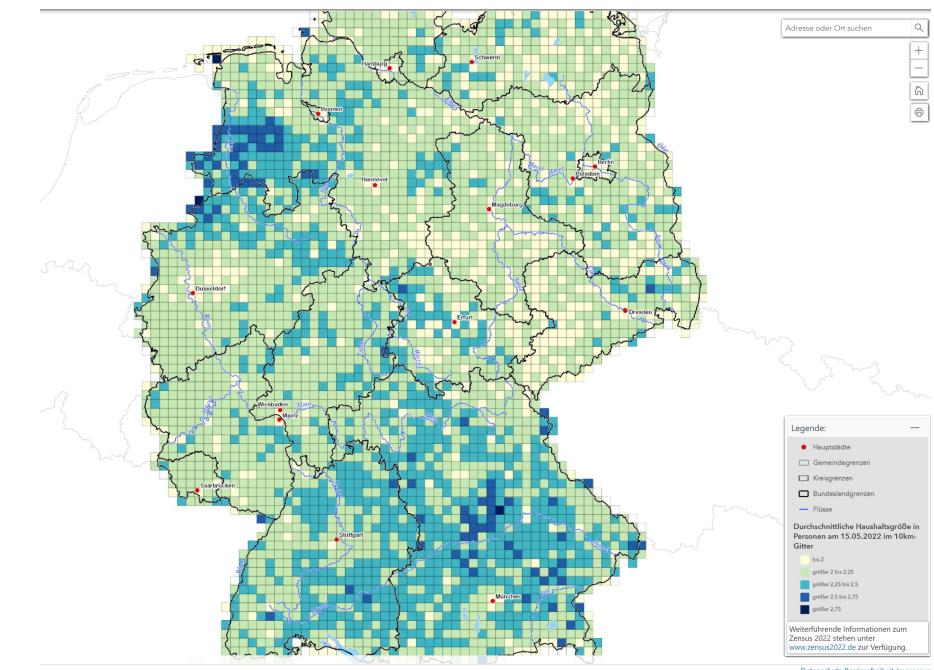
- Municipal data is linked with the administrative boundaries of the municipalities
 - The spatial mapping of these boundaries can be downloaded as polygons from the BKG (German cartographic agency) [as a shapefile \(.shp\)](#)
 - The shapefiles can be matched with the municipal dataset via the ID variables (AGS, ARS) e.g. with the *sf package* in R



Sources: <https://service.destatis.de/DE/karten/flaechenatlas.html>

Working with geospatial data

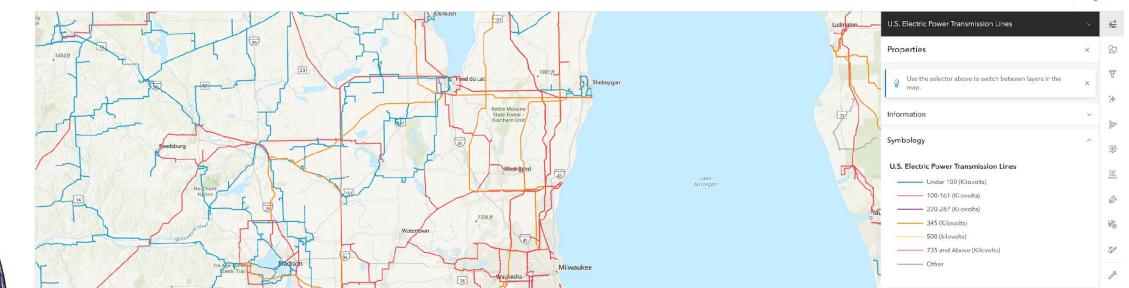
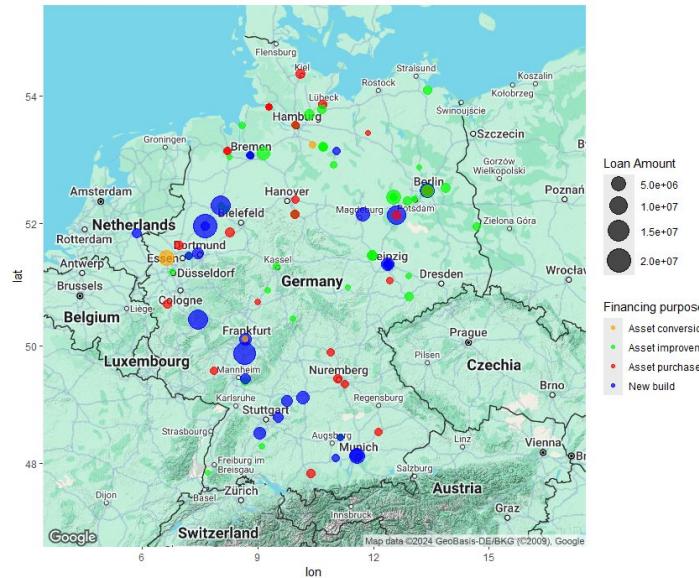
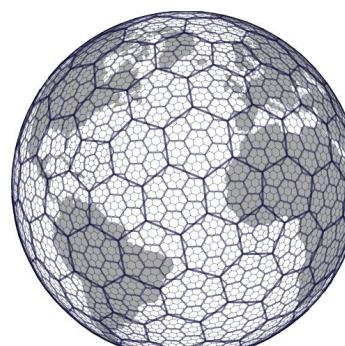
- Grid cells (also referred to as raster) all have the same spatial resolution
 - As such, they are more comparable than municipalities
 - However, grouping them by categories or spatial proximity might still be advisable
- When joining grid cells, ensure that coordinate projection systems (e.g. . EPSG:3857, WGS84 / EPSG:4326) match.
Our data is in EPSG 3035 coordinates



Sources: <https://atlas.zensus2022.de/>

Working with geospatial data

- Point data: Specific locations like a particular building have only one value for longitude and latitude resulting in a „point“
- Line data: A connection of points can represent a flow (e.g. a network or a river)
- Hexagonal: Similar to grid cells but makes neighbors equidistant



Sources:

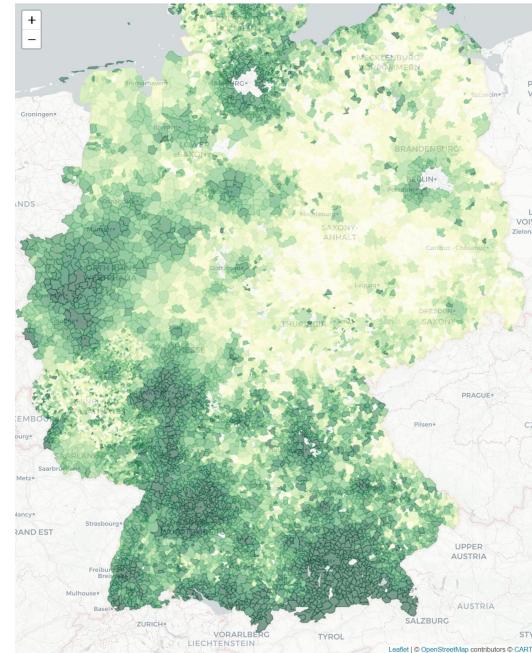
<https://www.arcgis.com/apps/mapviewer/index.html?layers=d4090758322c4d32a4cd002ffaa0aa>
12, own material

Working with geospatial data (methods)

- Visualization and segmentation of data based on geographic attributes
- Integrating geographical attributes into statistical techniques (e.g. Geographically weighted regression, spatial autocorrelation, spatial clustering). These address the issue that spatial observations are generally not independent of each other

Working with geospatial data (tools)

- R packages:
 - Working with shapefiles: sf, sp
 - Visualization: ggmap, tmap, leaflet (interactive)
 - Python packages (please ask consultants)
 - QGIS: Open source Geographic Information System softwares
 - Visuals and some analysis techniques



Sources: Own materials based on FDZ Ruhr RED V5 Immoscout ads

Linking with other data

- [BKG open data](#): administrative units, land cover points of interests, ...
- [Open Street Maps](#) (OSM): crowdsourced, many different layers (check out documentation), querying in R via Overpass API (osmdata package)
- Data sources mentioned in the variable description (Census 2011/2022, Urban Atlas, ...)
- [Regionalstatistik](#): Data on municipal and districts' land use (e.g. building permits), income, employment, etc.
- [Flächenatlas](#): Geospatial datasets from destatis
- Geocoding textual data: Location names and the attributes linked to the location can be „made spatial“ and mapped using a geocoder API (e.g. OSM Names or Google Maps Geocoder)

Coming up with questions: Ideation methods

Analytic approaches

Data-driven

e.g. predicting land use per person (or any other variable) based on a variety of variables such as vacancy rate, flat size, share of small households, ...

Theory-driven

e.g. “H1: In small municipalities(outside FUA zones), land use per person is positively correlated with the share of over 60 years old people reflecting the “empty nest effect”.”

Breadth-first

e.g. collect all variables that may influence the variable you seek to explain (e.g. land use)

Depth-first

e.g. deep dive into how one relationship manifests itself (e.g. vacancies ~ built surface increases/decreases) and explore it across different (regional) partitions and how it interacts with other attributes

Ideation methods

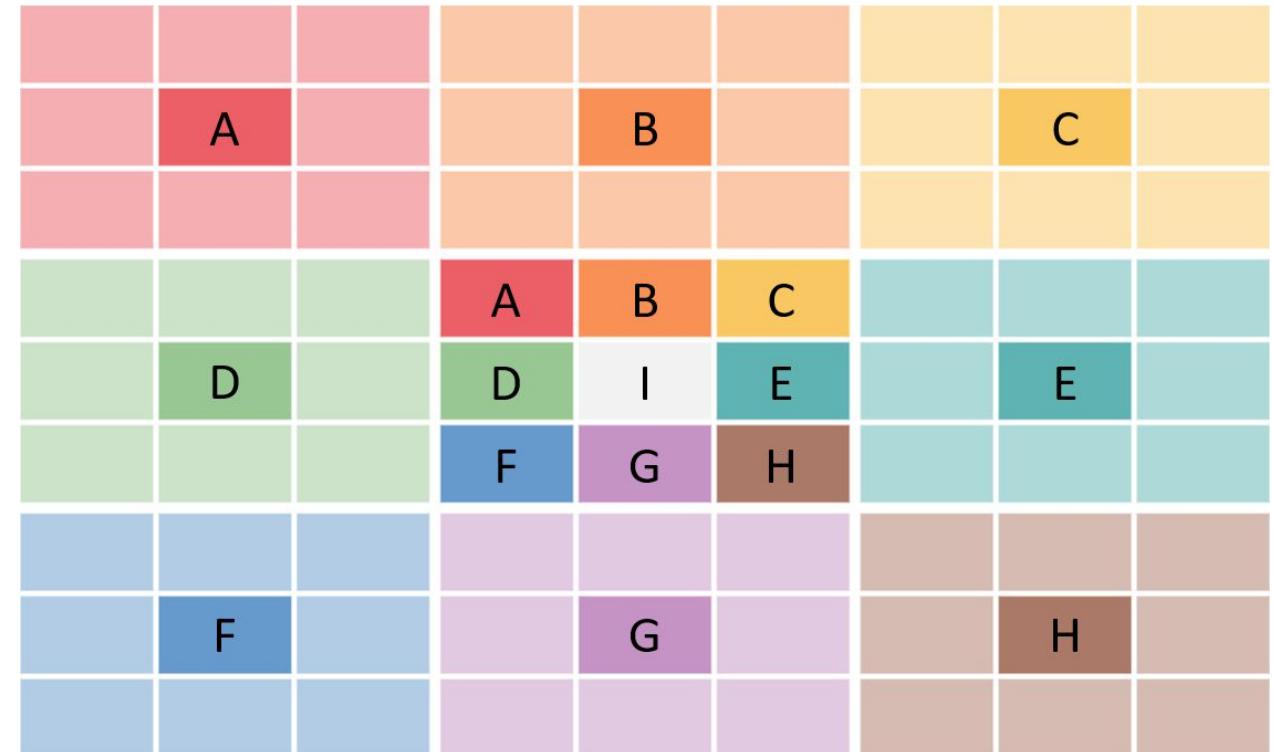
- 6-3-5 brainwriting method (in your case N-3-5)
 - N people, 3 ideas, 5 minutes
- Lotus diagram
 - Iterative refinement of the most promising ideas

N-3-5

Rounds	Idea 1	Idea 2	Idea 3
1	Identify municipalities with high land use	Which municipalities have the highest change rates in built surface?	How do municipalities in FUA commuting zones differ from those that are just outside the commuting zone?
2	Look at common socio-demographic characteristics and geographies of these municipalities
3	Are potential drivers of land take the same for all regions?		Don't judge at this point!
4	...		

Ideation Methods: Lotus Flower

- **Central Idea:** Start with a core concept or problem written in the center of a diagram, like the heart of a lotus flower (the I).
- **Petals of Ideas:** Surround the central idea with eight related ideas, questions or themes, each represented as a petal A-H excluding I).
- **Secondary Petals:** For each of the eight initial petals, generate eight more ideas, creating a second layer of petals. These secondary petals should be more specific and detailed, expanding on the initial petals.
- **Selection of ideas:** Select the preferred secondary approaches for further exploration



Ask an IIm ?

I have a dataset with German municipal level observations on land use, housing structures, age, population, ownership rates, vacancy rates, taxes. Some variables are present for multiple years while others are just point in time. What types of research questions could I explore with this data?

 With a dataset containing German municipal-level observations on various socio-economic indicators, you can explore a wide range of research questions. Here are some potential avenues for investigation:

1. Land Use and Urbanization Trends:

- How has land use changed over time in different municipalities?
- What are the patterns of urbanization and how do they correlate with population growth?

2. Housing Market Dynamics:

- How do housing structures (e.g., single-family homes vs. multi-family dwellings) vary across municipalities?
- What factors influence housing ownership rates and vacancy rates?

3. Demographic Analysis:

- How does the age distribution of the population vary across municipalities?
- What are the implications of an aging population on housing demand and land use?

4. Economic Impact of Taxes:

- How do different tax rates affect municipal economic indicators such as housing prices or vacancy rates?
- Is there a correlation between tax policies and population migration patterns?

5. Spatial Inequality:

- Are there significant disparities in housing and land use between urban and rural municipalities?
- How do socio-economic factors like income and education levels correlate with housing and land use patterns?

6. Policy Evaluation:

- What are the impacts of specific municipal policies (e.g., zoning laws, tax incentives) on land use and housing?
- How effective are municipal interventions aimed at reducing vacancy rates or promoting homeownership?

7. Environmental and Sustainability Considerations:

- How do land use patterns impact environmental sustainability in different municipalities?
- What are the trade-offs between urban development and preservation of green spaces?

8. Temporal Analysis:

- For variables available over multiple years, how have trends evolved, and what factors contribute to these changes?
- Are there cyclical patterns in housing market indicators, and how do they relate to broader economic cycles?

9. Comparative Analysis:

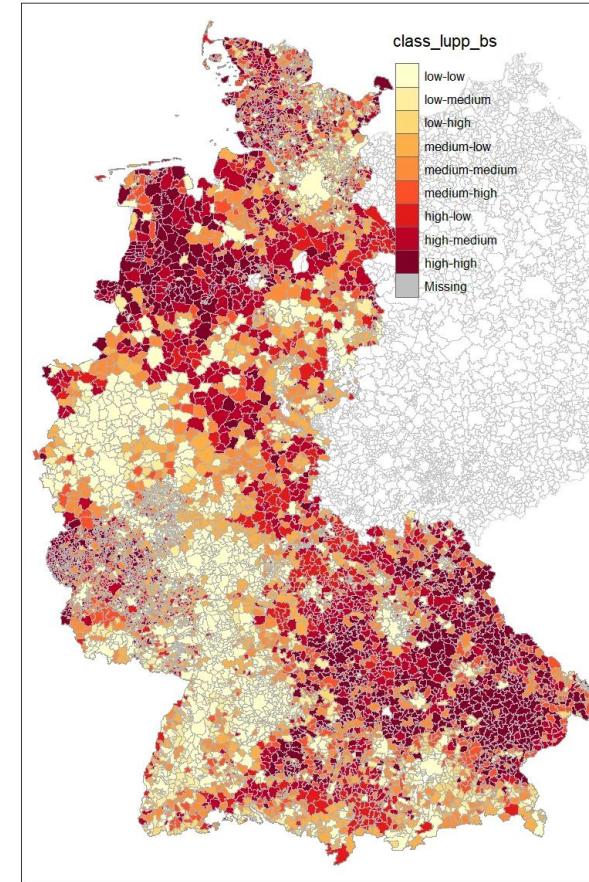
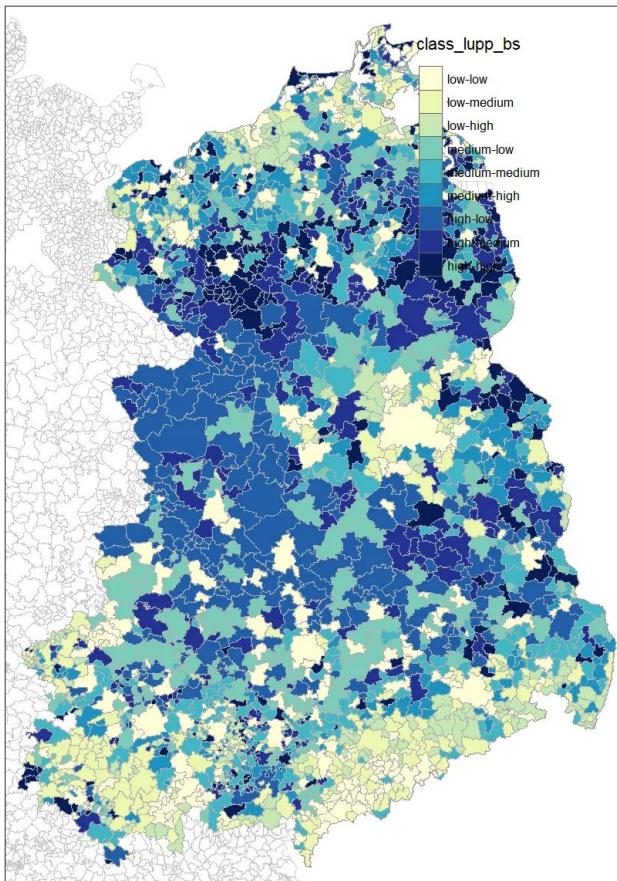
- How do municipalities compare in terms of housing affordability and accessibility?
- What can be learned from municipalities with successful housing and land use policies that could be applied elsewhere?

10. Predictive Modeling:

- Can future trends in housing demand and land use be predicted based on current data?
- What are the key drivers of change in the housing market and land use patterns?

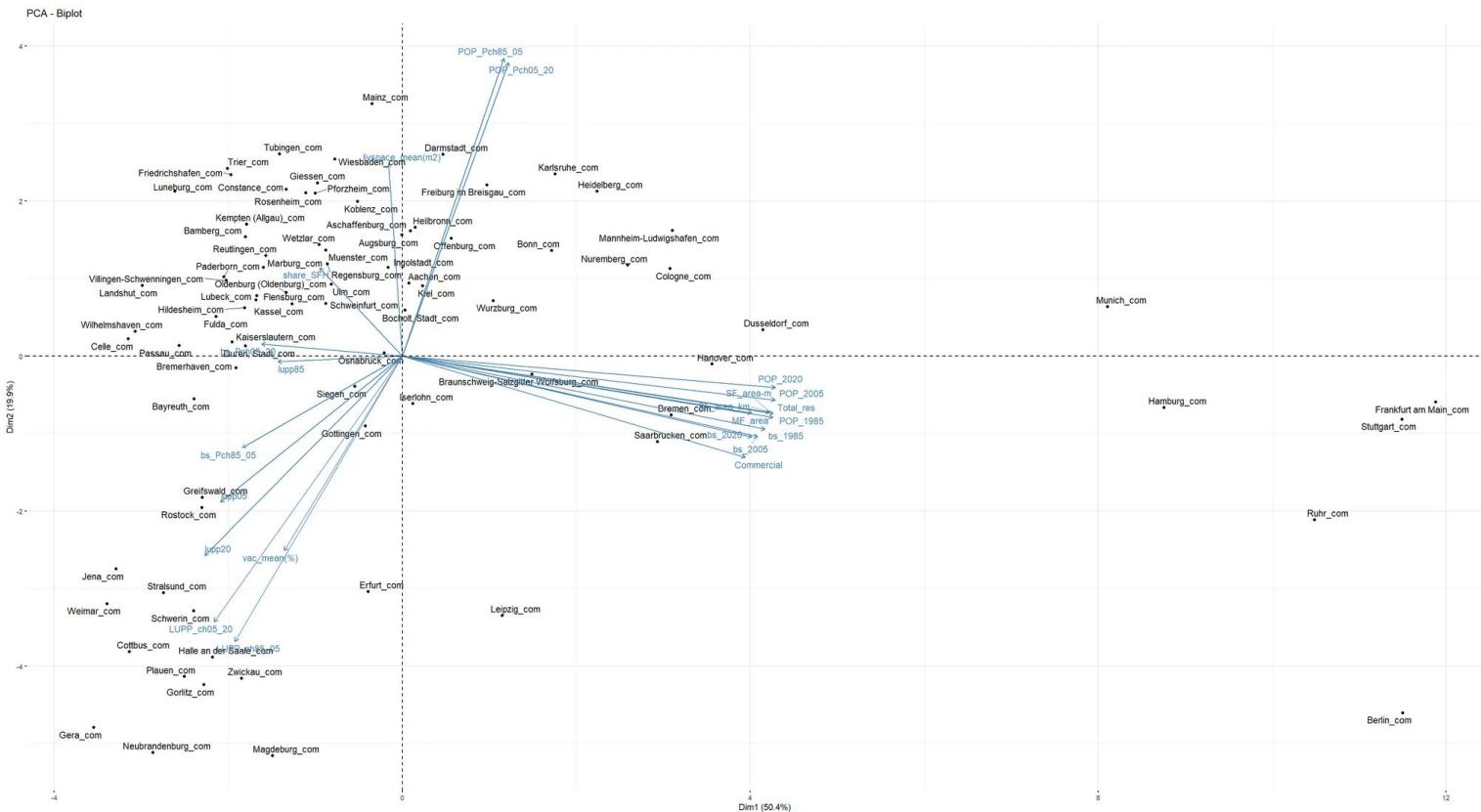
Example questions and approaches

Where are the least efficient municipalities in terms of LUPP and Built Surface Change in East and West Germany?



Example questions and approaches

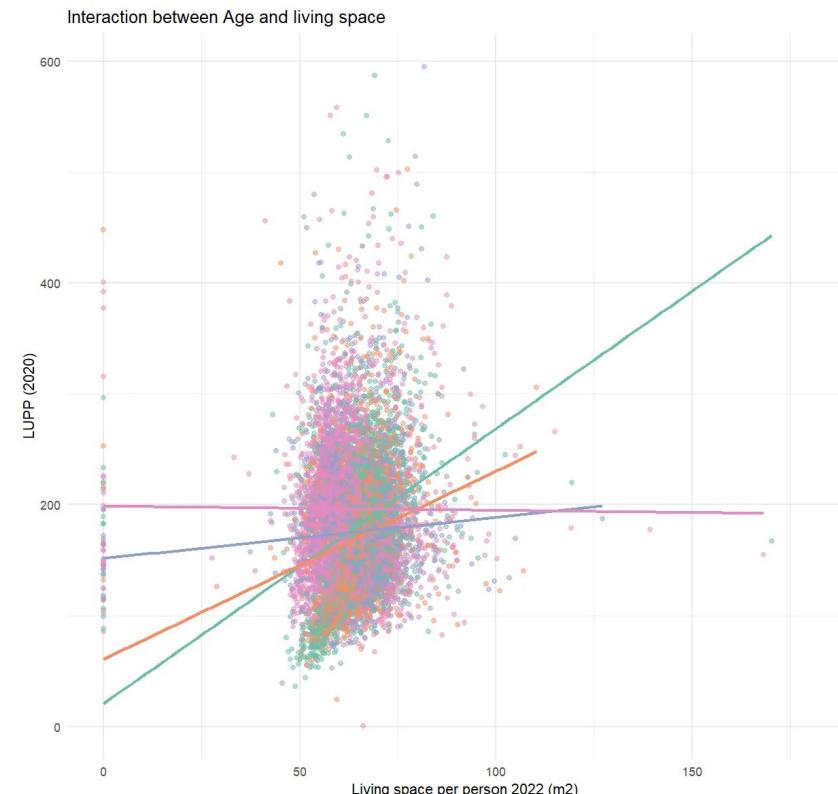
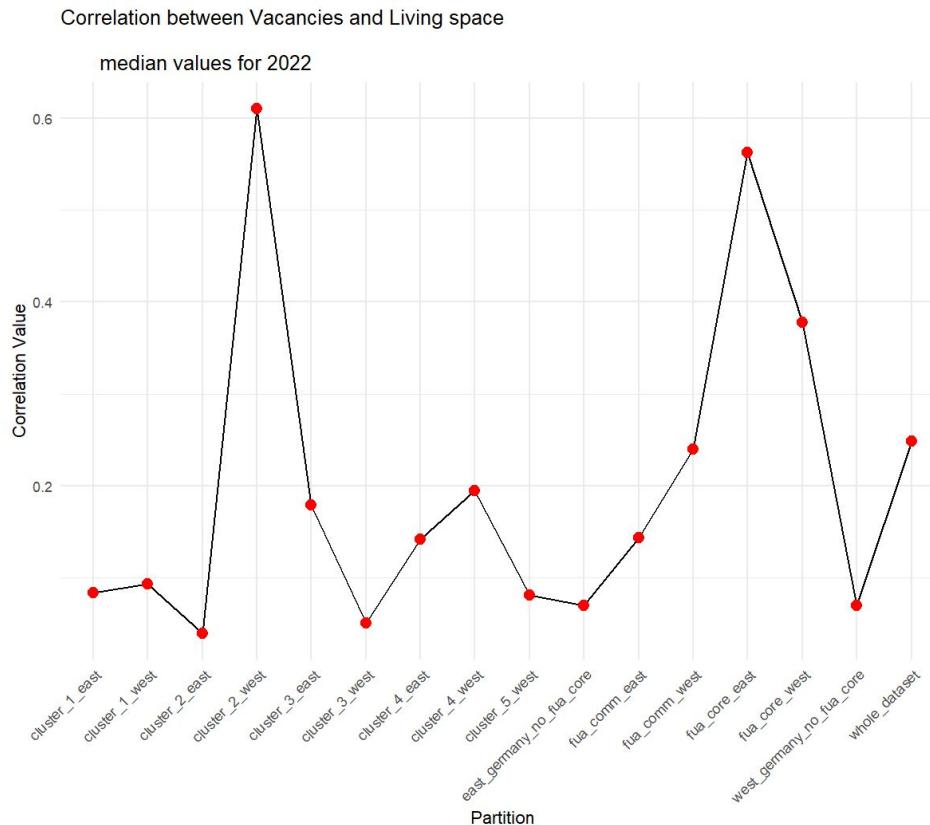
Group similar municipalities by reducing the dimensionality through Principal Component Analysis or Clustering



Example questions and approaches

Are there variations correlations across partitions of the dataset?

Which variables might be moderators?



Questions