Data Analysis using R

Introduction & Assignment

Sven Werenbeck-Ueding

18.11.2024

About This Course

- This course gives practical insights into conducting data analysis projects using R
- Covers data analysis process from beginning to end:
    1. Importing and cleaning raw data
    2. Exploratory data analysis and visualization
    3. Formulating the empirical model
    4. Communicating the results

All within R!

Learning Outcomes

By the end of the semester, you will...

... be able to conduct empirical projects on your own

... have a solid understanding of R and frequently used packages

... be able to use GitHub for version control of your code and collaboration

... be able to create dynamic, technical reports using Quarto

Course Structure

Lecture

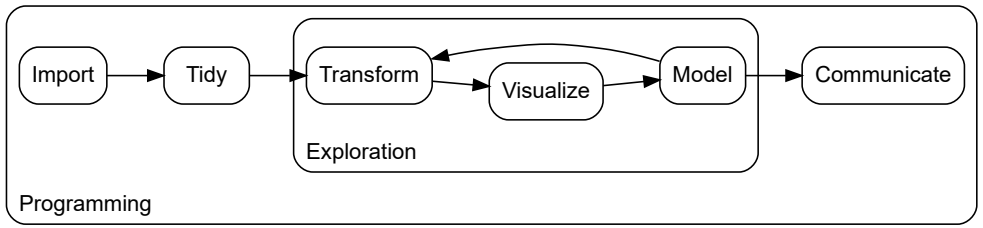
- Class times: Wednesday, 14:00 – 18:00, GD 03/354
- First half of the semester
- Each lecture covers a different topic
- Topics are shown using practical examples

Assignment

- Second half of the semester
- Working on an own data analysis project
  - In groups of 2 or 3 students
  - Presentation of outlines on **04.12.2024**
  - Submission of course work on **28.01.2025**
  - Presentation of results on **29.01.2025**
- Code repository and presentation count equally towards the final grade

Lecture: Agenda

Lecture	Topic
1	R Basics
2	Version Control
3	Programming
4	Importing
5	Data Wrangling
6	Visualization
7	Modeling
8	Reporting



Source: Wickham and Grolemund (2016)

Main Literature

Wickham, H. (2019). *Advanced R*. 2nd. Chapman & Hall/CRC. URL: http://adv-r.had.co.nz/.

Wickham, H. and G. Grolemund (2016). *R for data science. import, tidy, transform, visualize, and model data*. O'Reilly. URL: https://r4ds.had.co.nz/.

Contact Information

- Mail: sven.werenbeck-ueding@ruhr-uni-bochum.de
- Office: GD 03/367
- Office hours: Please make an appointment via email in advance

The Assignment

Housing Prices and School Quality

- Large number of studies find significant willingness-to-pay for living near better-performing schools, see e. g. the meta study by Machin (2011)
  - School performance usually measured through average test scores
  - (Elementary) schools in the neighborhood are a deciding factor for parents when buying a home
- Parents pay a mark-up to live near "better" schools with a good learning environment for their child
- Having good schools nearby can be understood as a local amenity

Reflected in higher prices for homes located near "better" schools

Valuing Amenities: Hedonic Prices

> [...] model of product differentiation based on the hedonic hypothesis that goods are valued for their utility-bearing attributes or characteristics.
>
> Rosen (1974)

Price for some complex good is defined by a function of its K characteristics, $p(x) = p(x_1, \ldots, x_k, \ldots, x_K)$, such that the price for characteristic k is the first-order derivative: $\frac{\partial p(x)}{\partial x_k} = p_k(z)$

Under general equilibrium assumptions, implicit prices of non-market goods (e. g. living near a good school) are revealed through product differentiation of complex goods (e. g. housing)

If two houses differ in one characteristic only, the difference in their prices is attributable to this characteristic

Hedonic Price Function

Hedonic price functions are usually estimated using a log-linear functional form:

$$\ln p_{itr} = S_{it}\beta + N_{it}\gamma + \tau_t + \rho_r + \varepsilon_{it}$$

- p_{itr}: Price of dwelling i at time t in region r
- S_{it}: Vector of structural characteristics of dwelling i at time t, e. g. the number of rooms
- N_{it}: Vector of neighborhood characteristics of dwelling i at time t
- τ_t: Time fixed effects for time t
- ρ_r: Region fixed effects for region r
- $\varepsilon_{it} \sim N(0, \sigma^2)$

$\hat{\beta}$ and $\hat{\gamma}$ give vectors of estimated implicit prices for structural and neighborhood characteristics under the c. p. condition

School Quality in NRW

- In North Rhine-Westphalia, schools are assigned a school social index reflecting the social composition of a school's student body
  - A bad score indicates that a school is confronted with severe challenges
  - Might indicate a less favorable learning environment for a child
  - Schools with bad scores are allocated more resources
  - Bad learning environment probably outweighs the additional resources
- Similar to the example with average test scores, parents would likely choose to buy a home near schools with a better school social index

Neighborhoods with "better" schools may attract more parents, resulting in higher prices for houses/apartments

School Social Index in NRW

- School-specific social index for general public education schools (primary schools, secondary schools, i. e. *Hauptschule*, *Realschule* and *Sekundarschule*, comprehensive schools and grammar schools)
- The index ranging from 1 (good) to 9 (bad) reflects the social composition of a school's student body and is taken into account in the distribution of resources
- The social composition of the pupils in the schools is mapped via the following four indicators:
    1. Children and youth poverty
    2. Share of pupils with predominantly Non-German speaking households
    3. Share of pupils with migration background (first-generation immigrants)
    4. Share of pupils with special needs (focus: learning, emotional and social development, speaking)

Note: More information can be found here (unfortunately, only in German… 🙄)

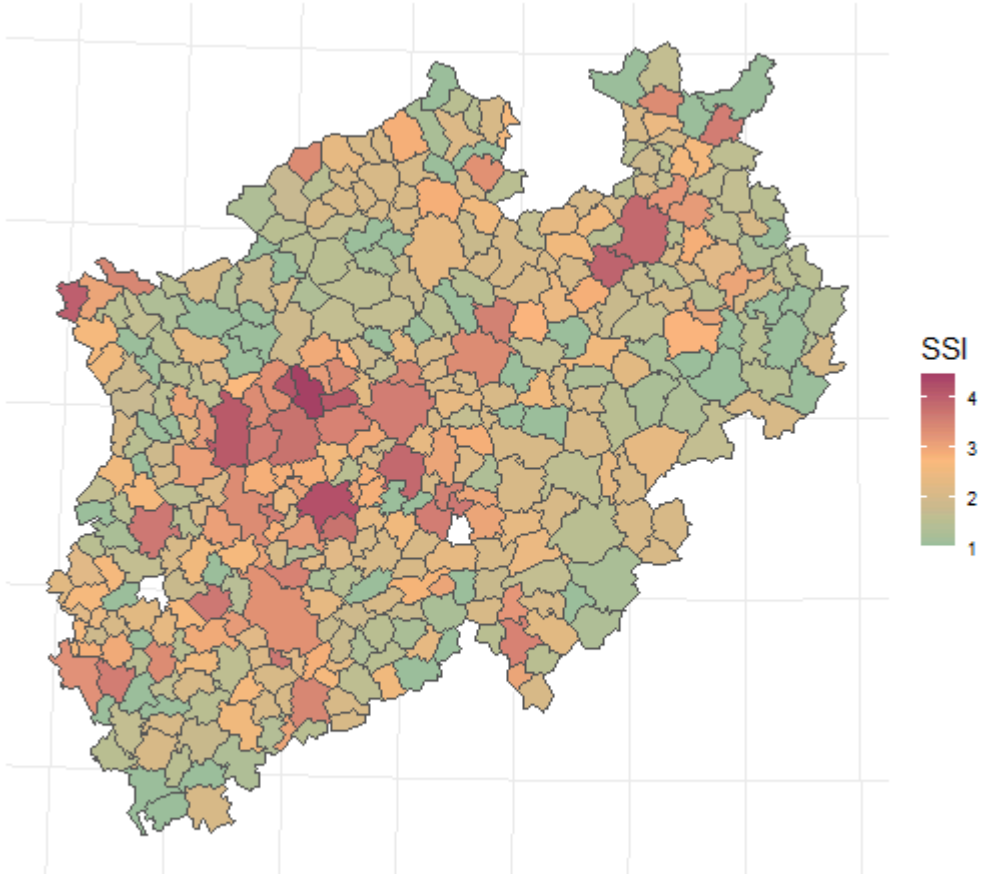
School Social Index in NRW

	School social index										
School type	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**w/o index**	**Sum**
Primary education											
Elementary school	644	823	529	300	175	145	70	14	4	8	2712
Secondary education											
Secondary school (Hauptschule)	1	7	25	40	35	32	27	7	4	1	179
Secondary school (Realschule)	48	95	104	53	25	3	3	-	-	1	332
Secondary school	4	44	36	11	5	2	-	-	-	5	107
Comprehensive school	30	129	91	39	9	7	2	-	-	11	318
Grammar school	260	186	45	10	1	1	1	-	-	-	504

Note:

School social indexes were calculated on the basis of the Official School Data 2018/2019. The data was taken from: https://www.schulministerium.nrw/sozialindex (last accessed on 03.10.2023).

School Social Index in NRW



Note: SSI were averaged over elementary schools of municipalities.

Your Task

- Develop a specific and narrow research question revolving around housing prices and school quality in groups of 2 to 3 students
- Prepare the provided data to your needs
  - Online advertisements from *Immobilienscout24* containing houses and apartments offered for sale and apartments offered for rent in 2022
  - School social index from NRW's Ministry of Schools and Education
  - Further data you consider relevant for your research question (the database of the Federal Statistical Office of Germany or the regional database INKAR)
- Formulate an empirical strategy designed to answer your research question
- Develop your code and collaborate in a GitHub repository
- Present your results in class at the end of the semester
- Code repository and presentation count equally towards the final grade

Timeline

Date	
16.10.2024 – 30.10.2024	Register via Flexnow Form groups of 2 to 3 students (use the Moodle forum)
today – 31.10.2024	Register on sciebo: https://hochschulcloud.nrw/ Send an email to sven.werenbeck-ueding@rub.de containing the members of your group with name, student ID and the email with which they registered on sciebo
04.11.2024	Copy the data from the shared sciebo folder to your own folder in sciebo and start working on your project by creating a GitHub repository dedicated to your project Add the GitHub user "Solaire-patch" (that's me) as a collaborator to your GitHub repository
04.12.2025	Present an outline of your project
28.01.2025	Create a feature branch labelled "submission" to submit your repository and also submit your repository as a ZIP archive via email to sven.werenbeck-ueding@rub.de
29.01.2025	Present your results in class

Presentation of Project Outlines

- **5 to 10 min.** presentation of your research idea
  - What is your research question?
  - How do you motivate your research question?
  - What are key insights from related literature?
- Provide an overview of your empirical strategy
  - What data will you use and how are you processing the data?
  - What is your empirical model?
  - Are there issues with your empirical strategy?
- Presentation on **04.12.2024**

Code Repository

- GitHub repository containing **all** code and results
- Collaborate in the repository with your group members
- Use the features provided by GitHub:
  - Split your analysis in reasonable subtasks using GitHub issues
  - Work on feature branches to complete the tasks
  - Make use of commits and commit messages
  - Use pull requests to review each other's code
- Follow the guidelines stated in the lecture(s)

Do not push the provided data to your code repository!

Presentation

- **15 min. presentation** followed by a **5 min. discussion**
- Every group member has to contribute to the presentation to equal parts
- Has to be created using Quarto and submitted as PDF
- Should give insights on
  - Motivation, research question and contribution to the existing literature
  - Data used and empirical strategy
  - Main findings
  - Limitations of your approach

General Remarks on the Presentation

Introduction

- Introduce and motivate your topic
- What is your research question?
- What are the main findings of the related literature and how does your research contribute to it?

Data and Empirical Strategy

- What data do you use and how did you pre-process it? Were some observations omitted? If so, why?
- Which empirical strategy do you use to answer your research question? And why this approach?

Results

- What are the main findings?
- Are there heterogeneous effects?

Conclusion

- Summarize your findings and set them in the context of the existing literature
- State the limitations of your approach

Submission

Deadline: 28.01.2025

Code Repository

- GitHub repository for collaboration with your group members and containing **all** code and results of your group work (including the presentation as PDF)
- Add "swerenbeck" (that's me) as a collaborator
- Create a feature branch labelled "submission" which will be graded

Email

- One email per group to sven.werenbeck-ueding@rub.de
- Attach the whole repository from the feature branch "submission" as a ZIP archive and the presentation as PDF
- To create a ZIP archive right-click with your mouse on the repository folder on your local machine and select "In ZIP-Datei komprimieren" / "Compress to ZIP file"

The Data

Housing Data

- The *Research Data Center Ruhr* (FDZ Ruhr) publishes on a regular basis all residential properties advertised on the internet platform *ImmobilienScout24* that are offered for sale or rent
- Datasets provide detailed information on housing attributes and location of these properties
- You will be provided access to a campus file containing all advertisements from 2022

Citation:

RWI and ImmobilienScout24 (2023). RWI Real Estate Data - Campus File Cross-Section. Version: 1. RWI – Leibniz Institute for Economic Research. Dataset. doi.org/10.7807/immo:red:cross:v4

School Data

The provided data on schools contains the following data sets:

- `school_data.csv`: Basic information on all schools in NRW
  - School type
  - Legal form
  - ZIP code area
  - District
  - ...
- `2022_social_index.csv`: Social school indexes per school for the year 2022
- `number_pupils.xlsx`: Number of pupils/students per school
- `keys.xlsx`: Labels for the keys `legal_form`, `district`, `school_operation` and `school_type` in the `school_data.csv` data set
- `distance_to_schools.csv`: Distance between a square kilometer grid cell's centroid and the first as well as second nearest school (for each school type)
- `plz_AGS.csv`: Mapping between ZIP codes and the AGS (*Allgemeinder Gemeindeschlüssel*, municipality code)

Note: The data was taken from the North Rhine-Westphalia Ministry of Schools and Education (https://www.schulministerium.nrw/open-data, last accessed on 03.10.2023)

Regional Data

The data set `region_data.csv` contains demographics on municipality-level in NRW (*Gemeinde*):

- `munic_name`, `AGS`: Name and code of the municipality
- `population`: Population count
- `work_age_population`: Working age population count (16 to 65 years old)
- `unemployed`: Number of unemployed
- `low_income_hh`, `middle_income_hh`, `high_income_hh`: Share of low (< 1.500€ per month), middle (between 1.500€ and 3.600€ per month) and high (> 3.600€ per month) income households
- `population_density`: Population density (inhabitants per km^2)
- `migrants`: Number of migrants
- `region_type`: Region type, i. e. urban (metropolitan), urban (regiopolitan), rural (close to city region) or rural (peripheral)

Note: The number of migrants was taken from the regional database of the Federal Statistical Office and is based on the 2011 census and the rest of the data was obtained from INKAR and is from 31.12.2020

Data Access

After the registration period and when all groups are formed, the data will be provided to you via sciebo:

- Cloud service with 30GB storage
- Register via your RUB-mail (or any other mail from a higher education institution in NRW)
- A folder containing the data will be shared with you

Do not store the data anywhere else than sciebo!

Some Limitations

Not controlling for socio-economic indicators will confound your estimates for the implicit prices of school quality since both are correlated

⚠ Regions with high unemployment, share of migrants etc. will have higher SSI scores

⚠ These indicators are not available to you on the lowest regional aggregation level of schools (ZIP code areas)

⚠ You may not be able to control for the location of a dwelling inside a region

References

Machin, S. (2011). "Houses and schools: Valuation of school quality through the housing market". In: *Labour Economics* 18 (6), pp. 723-729. DOI: https://doi.org/10.1016/j.labeco.2011.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0927537111000601.

Rosen, S. (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition". In: *Journal of Political Economy* 82.1, pp. 34-55. URL: http://www.jstor.org/stable/1830899.

Wickham, H. (2019). *Advanced R*. 2nd. Chapman & Hall/CRC. URL: http://adv-r.had.co.nz/.

Wickham, H. and G. Grolemund (2016). *R for data science. import, tidy, transform, visualize, and model data*. O'Reilly. URL: https://r4ds.had.co.nz/.