Lecture 3

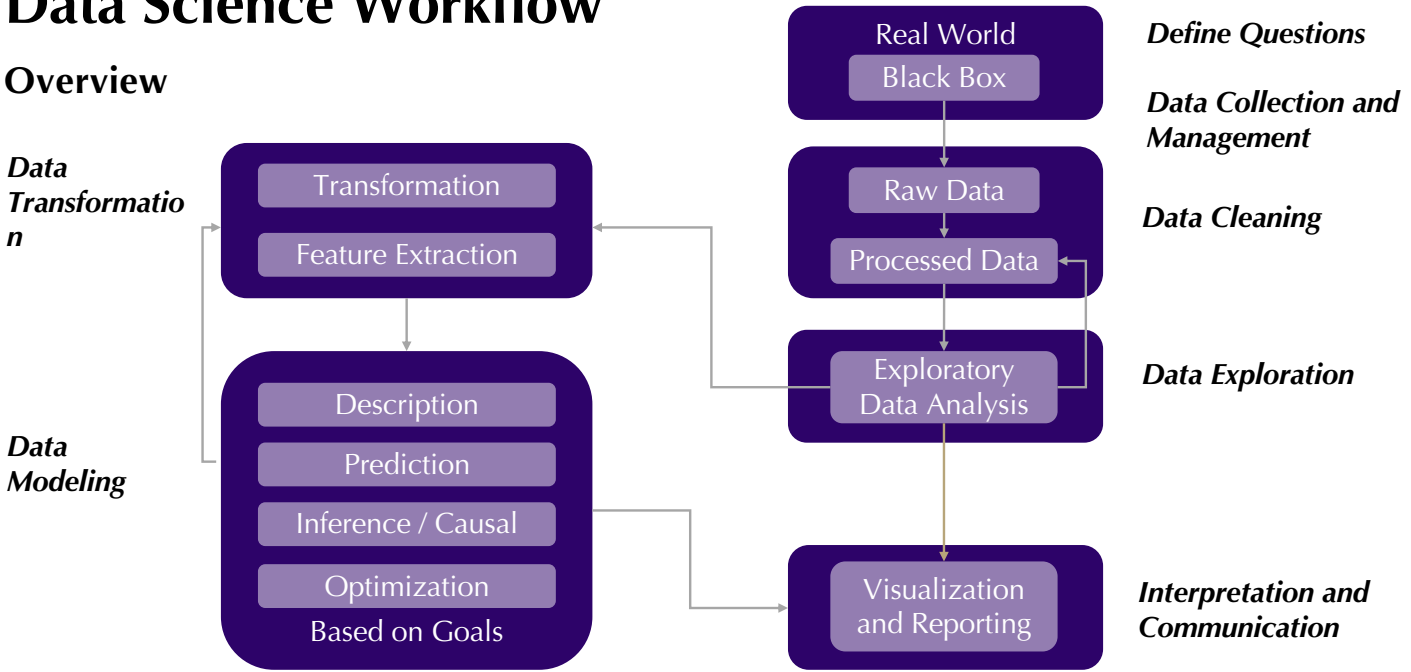
Data Science Workflows and Cases

Haoyu Yue / yohaoyu@washington.edu
Ph.D. Student, Interdisciplinary Urban Design and Planning
University of Washington

RE 519 Real Estate Data Analytics and Visualization
Course Website: www.yuehaoyu.com/data-analytics-visualization/
Autumn 2025

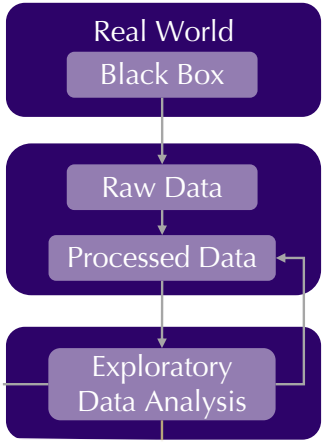
Data Science Workflow

Overview



Data Science Workflow

From Real World to Data Representation



Define Questions

Data Collection and Management

Data Cleaning

Data Exploration

Define the questions and some key issues, such as:

- Unit of analysis (household, property, region, etc.)
- Temporal and spatial extent
- Variables, including explanatory and outcome variables

Get your data and understand it:

- The source and quality of datasets
- Population or samples? Data types? Measurements?
- Data licensing and privacy/ethics issues

Start to process the datasets:

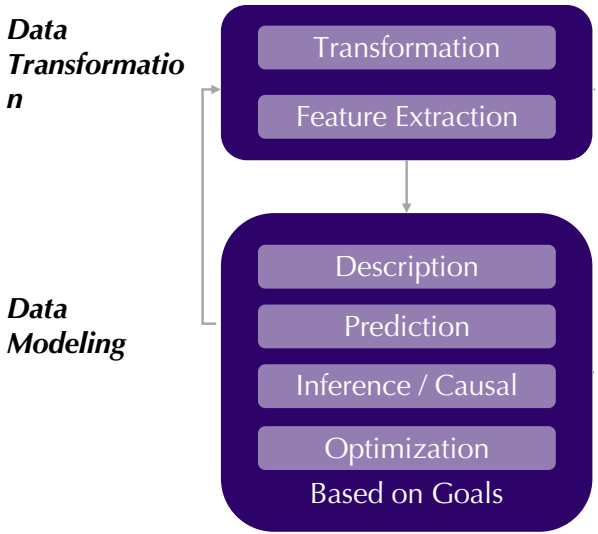
- All kinds of errors in the data, such as missing values, outliers, data replication, and different spatial boundaries
- All kinds of inconsistencies in the data, such as different units, date formats, and naming styles
- Tidy data principle (we will talk about this in the lab)

Conduct exploratory data analysis:

- Explore descriptive statistics, the distribution of variables, and the relationship among variables
- Refine research questions and clean the data again if needed

Data Science Workflow

Data Modeling



Data transformation before modeling:

- You need to have a sense of which model to use.
- Do some numeric transformation if needed: log transformation, standardized
- Construct new variables if needed, such as GDP per capita
- Select the most important variables using domain knowledge, correlation, Principal Component Analysis (PCA), etc.

Start data modeling (we classify based on purposes here):

- **Description** (especially if you have data for the population)
  - To understand the patterns, we can use visualization, clustering, and PCA etc.
- **Prediction** (we do not care about the relationship, but the accuracy)
  - Regression, machine learning, and deep learning, etc.
- **Inference / Causal** (we care about their relationship!)
  - Regression, causal inference, and hypothesis testing
  - **Correlation is not a causal relationship!**
- **Optimization** (find the best solution)
  - Simulation, linear programming, etc..

Data Science Workflow

Visualization and Communication

When you have some conclusions after data analysis:

- You need to explain and report your results to audiences.
  - Who is your audience?
  - What is the key message/takeaway?
- Make nice and effective visualizations to support your message
- Explain the results and talk about the implications
- Maybe suggest actionable steps
- Acknowledge the limitations of your analysis

Interpretation and Communication

The workflow we discussed is the ideal picture of data analysis and a simplified one.

In reality, the process is messy, iterative, and full of trade-offs.

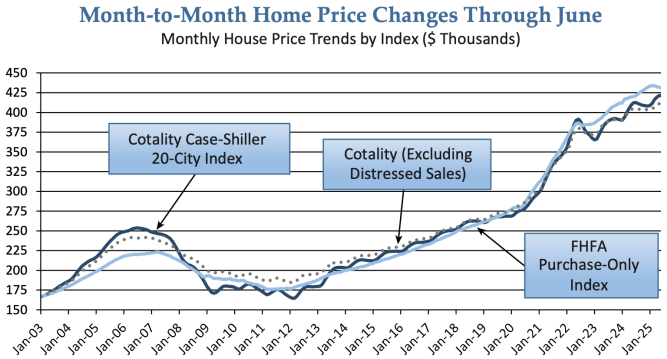
The real skill is to adapt, revise, and still deliver credible and useful insights.

Examples of Data Science in Real Estate

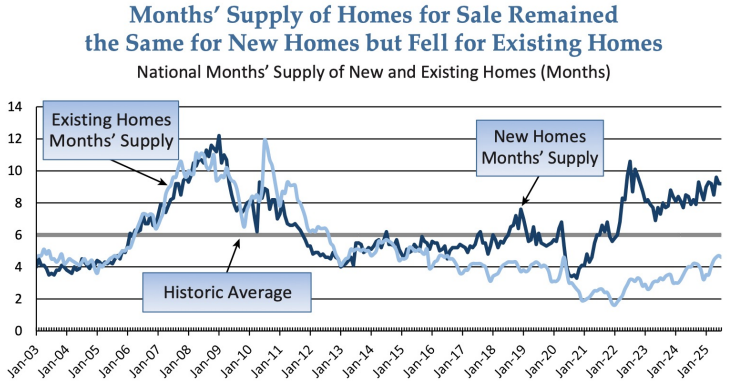
Descriptive – Predictive – Diagnostic – Prescriptive

Inferential / Causal

WHAT HAPPENED?



Sources: Standard & Poor's, Federal Housing Finance Agency, Cotality (formerly CoreLogic), and HUD.
See Note 1, Sources and Methodology.



Sources: Census Bureau, National Association of REALTORS®, and HUD.

Office of Policy Development and Research. (2025). Housing Market Indicators Monthly Update August 2025.
https://www.huduser.gov/portal/ushmc/hmi-update.html

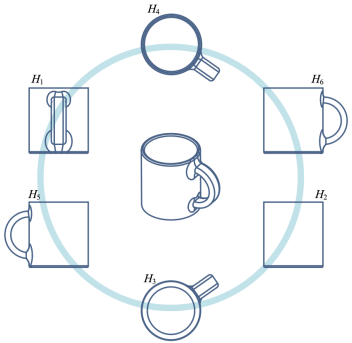
Examples of Data Science in Real Estate

WHAT HAPPENED?

Descriptive – Predictive – Diagnostic (Inferential / Causal) – Prescriptive

17 Variables → Principal Component Analysis (PCA) → Reduced to 5 PCs

Market size and dynamics
Population [1]
Population growth (%) [1]
Housing stock (no. of flats) [1]
Vacancy rate (%) [1]
Price level and dynamics
Rents (€/sqm) [2]
Rental growth (%) [2]
Purchase prices condominiums (€/sqm) [2]
Purchase price growth (%) [2]
Gross initial yield (%) [2]
Socioeconomic indicators and dynamics
Purchasing power per household (€) [3]
Purchasing power growth (%) [3]
Unemployment (%) [4]
Rent affordability (%) [1, 2, 3]
Price affordability [1, 2, 3]
Ownership rate (%) [1]
Demographics
Age cohort: 18–35 years (%) [1]
Change of age cohort: 18–35 years (%) [1]



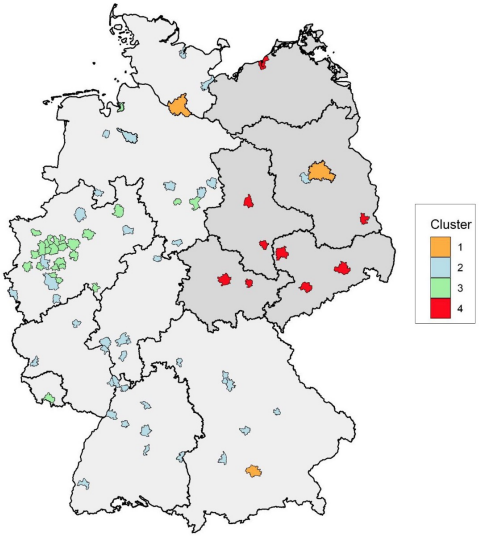
PCA Illustration.
Source: Visualize ML: https://github.com/Visualize-ML

↓

K-means Clustering

Example Conclusion:

German residential markets can best be segmented into four groups.



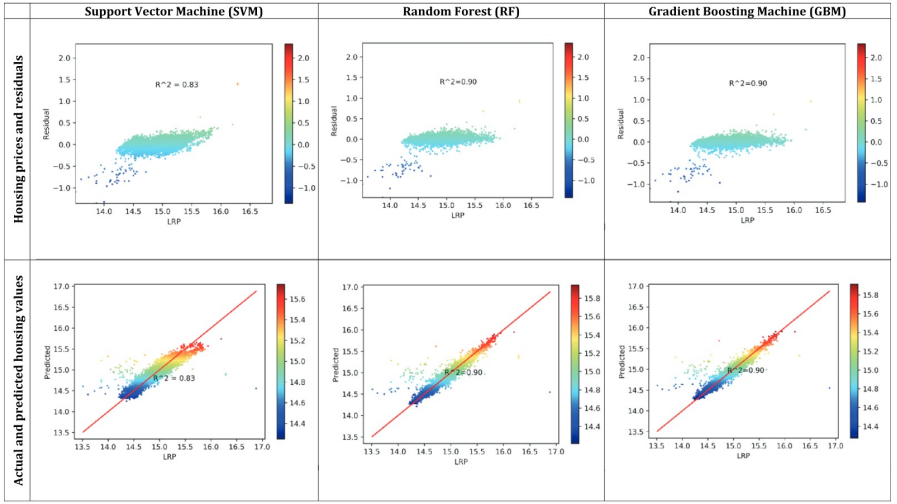
Note: The five new federal states (former GDR) are depicted in dark grey

Wiersma, S., Just, T., & Heinrich, M. (2022). **Segmenting German Housing Markets Using Principal Component and Cluster Analyses**. International Journal of Housing Markets and Analysis, 15(3), 548–578.
https://doi.org/10.1108/IJHMA-01-2021-0006

Examples of Data Science in Real Estate

WHAT WILL HAPPEN?

Descriptive – **Predictive** – Diagnostic – Prescriptive
Inferential / Causal



Mostly based on machine learning algorithms, the most important performance indicator is **prediction accuracy.** Because we care about the prediction power of the model on future (unseen) data.

Example Conclusion:

Each method can achieve an accuracy at XX levels in property price prediction.

Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2021). **Predicting Property Prices with Machine Learning Algorithms**. Journal of Property Research, 38(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Examples of Data Science in Real Estate

WHY HAPPENED?

Descriptive – Predictive – **Diagnostic** – Prescriptive

Inferential / Causal

Hedonic price model coefficients		Coefficients	Significance (p)
β_0	**Bracket** (whether or not a pair of sales for a property occurred on either side of one or more flood events within a postcode)	14.0%	< 0.001
α_0	**Bracket × Risk High**	–6.6%	0.051
α_1	**Bracket × Flood Zone**	–9.4%	0.092
α_2	**Bracket × Flood Zone × Risk High**	–21.8%	< 0.001
α_3	**Bracket × Flood Zone × Flood History**	19.0%	0.034
α_4	**Bracket × Flood Zone × Flood History**	–12.5%	< 0.001
α_5	**Bracket × Flood Zone × House Type**	9.1%	0.215
α_6	**Bracket × Flood Zone × Recovery**	24.2%	< 0.001
γ_0	**Year of First Sale**	–6.5%	< 0.001
γ_1	**Year of Second Sale**	3.9%	< 0.001

Thompson, J. J., Wilby, R. L., Hillier, J. K., Connell, R., & Saville, G. R. (2023). **Climate Gentrification: Valuing Perceived Climate Risks in Property Prices.** Annals of the American Association of Geographers, 113(5), 1092–1111. https://doi.org/10.1080/24694452.2022.2156318

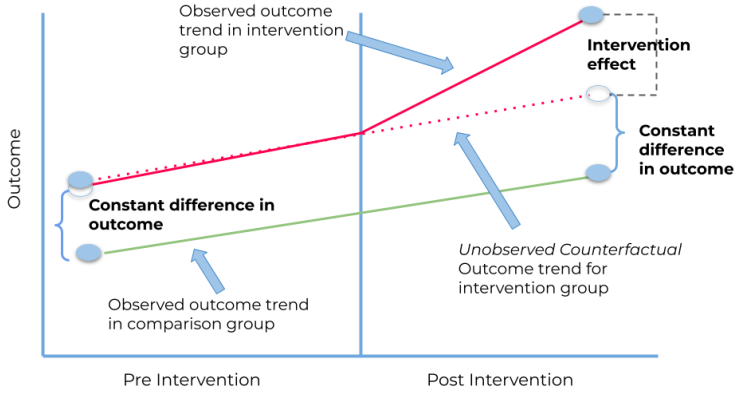
Example Conclusion:
The statistically significant relationship (**associations**) between housing prices and other variables.

Examples of Data Science in Real Estate

WHY HAPPENED?

Descriptive – Predictive – **Diagnostic** – Prescriptive

Inferential / **Causal**



DID. Source: https://medium.com/bukalapak-data/difference-in-differences-8c925e691fff

Cao, J., Huang, B., & Lai, R. N. (2015). **On the Effectiveness of Housing Purchase Restriction Policy in China: A Difference in Difference Approach**. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2584275

Whether one intervention **causes** the changes in the outcome.

Example Conclusion:

Housing purchase restriction policy in China **triggered** a substantial decline in the property price and transaction volume. The policy had no measurable **effects** on the nationwide construction boom.

Examples of Data Science in Real Estate

Descriptive – Predictive – Diagnostic Inferential / Causal – **Prescriptive**

WHAT SHOULD DO?

in Millions		Inflex 30	Flex 30	Flex 40	Flex 50	Inflex 60
Expected NPV		($191)	$14	$9	$4	$14
Median		($251)	($145)	($137)	($128)	($95)
Mode		($300)	($500)	($500)	($300)	($300)
Std Deviation		$371	$710	$714	$719	$731
Percentiles						
Value	5%	($686)	($792)	($847)	($901)	($962)
At	10%	($604)	($708)	($742)	($770)	($808)
Risk	25%	($455)	($543)	($535)	($527)	($502)
Median	50%	($251)	($145)	($137)	($128)	($95)
Value	75%	$7	$400	$397	$388	$414
At	90%	$295	$979	$963	$948	$977
Gain	95%	$509	$1,383	$1,372	$1,349	$1,380

Figure 18: 2 WTC Financial Model Results

The flexible 30 floor design and the inflexible 60 floor design outperform the other buildings. Note that the flexible 30 floor design has the lowest potential losses, yet maintains good gains when the economy is good.

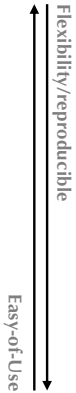
Leung, K. C.-K. (2014). **Beyond DCF Analysis in Real Estate Financial Modeling: Probabilistic Evaluation of Real Estate Ventures** [Massachusetts Institute of Technology]. https://dspace.mit.edu/bitstream/handle/1721.1/87612/879666642-MIT.pdf;sequence=2

A decision framework and prescriptive modeling steps to guide practitioners on how to manage uncertainty and make better ex ante investment decisions.

Example Conclusion:
Under a particular condition/scenario, the expected return is XXX, while the associated uncertainty is XXX.

Tools for Data Analysis

Some Example Tools



Basic

- Excel/Google Sheets – they help understand data and some data cleaning

Coding-lite

- SPSS/Stata – widely used in social science
- ArcGIS/QGIS – for geospatial analysis

Programming-based

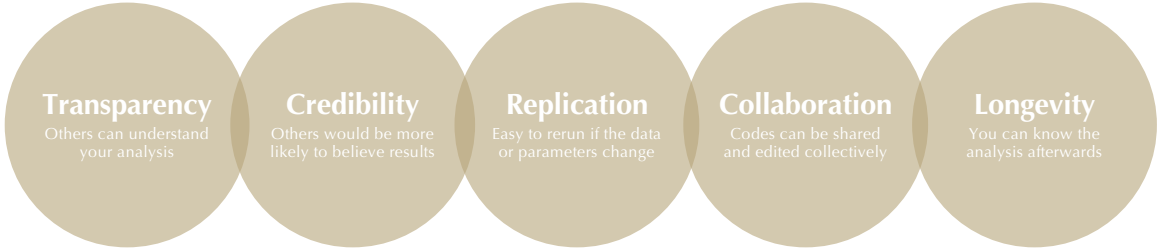
- R/Python – general-purpose language for regression, machine learning, visualization, and more
- SQL – for managing/querying large datasets

Urban/Real Estate Tools

- CoStar - a platform for commercial real estate information and analytics (please let me know if you want to get access to the platform this quarter)
- UrbanSim – urban simulation and planning, developed by a previous UW professor, Paul Waddell

Tools for Data Analysis

Value of Reproducibility and Open



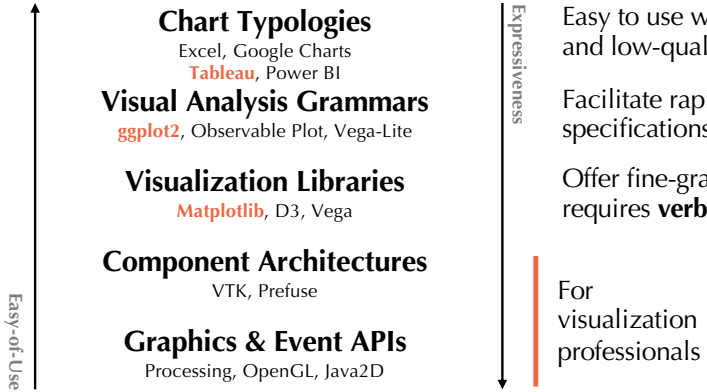
How to Achieve Open and Reproducible Data Science

- Use Programming to Process Data like Python or R
- Use Expressive Names for Files and Directories to Organize Your Work
- Use Findable, Accessible, Interoperable, and Re-usable (Wilkinson et al. 2016) Data
- Protect Your Raw Data
- Use Version Control (Git and GitHub) and Share Your Code
- Document Your Workflows
- Design Workflows That Can Be Easily Recreated

Source: What Is Open Reproducible Science. https://earthdatascience.org/courses/intro-to-earth-data-science/open-reproducible-science/get-started-open-reproducible-science/

Tools for Data Visualization

Out-of-the-Box and Programming Visualization Tools



Source: Visualization Tools, CSE 512 by Jeffrey Heer; https://courses.cs.washington.edu/courses/cse512/25sp/

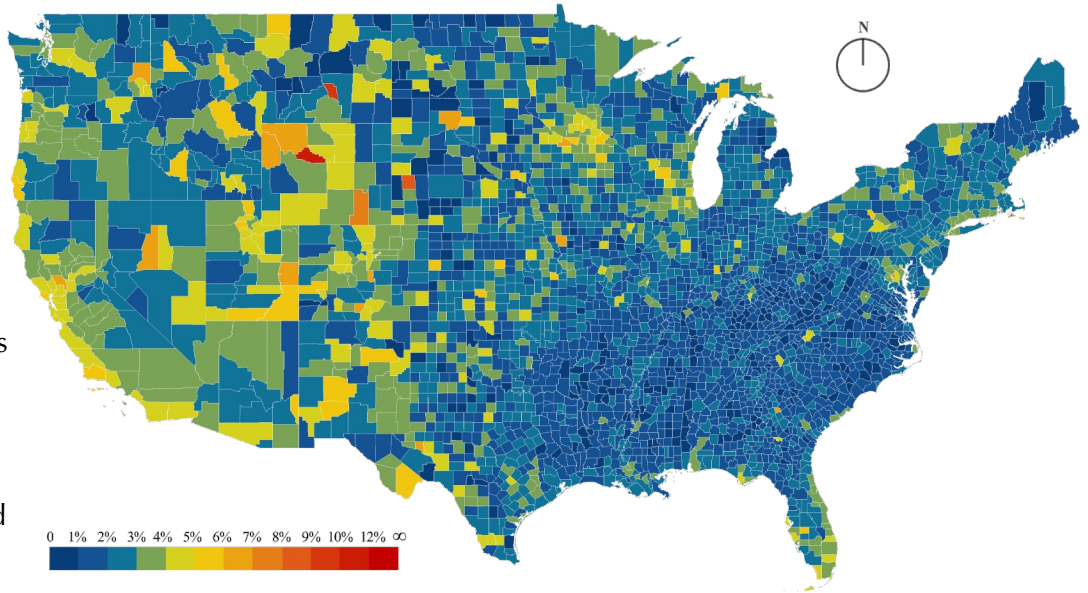
Tools for Data Visualization

Mapping Tools

ArcGIS/QGIS
- Lots of spatial analysis and mapping tools
- No coding required
- RE 497/597, URBAN 504

Mapbox
- Online mapping platform
- Pay for advanced functions

Programming Approach
- We will cover some mapping techniques in R
- Hard to navigate maps and not intuitive



Other Tools for Data Analysis/Visualization

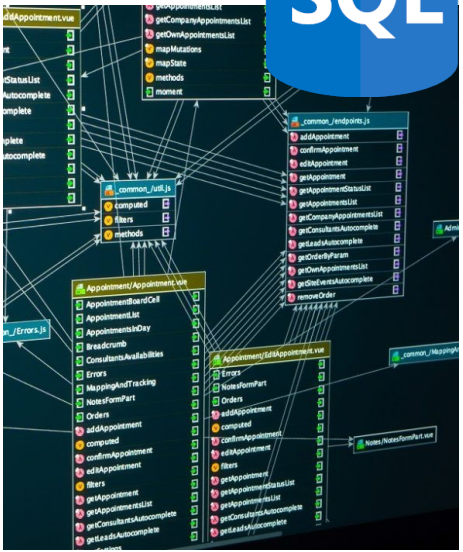
Microsoft PowerPoint

Adobe Photoshop

Adobe Illustrator

Git + GitHub

SQL Platforms

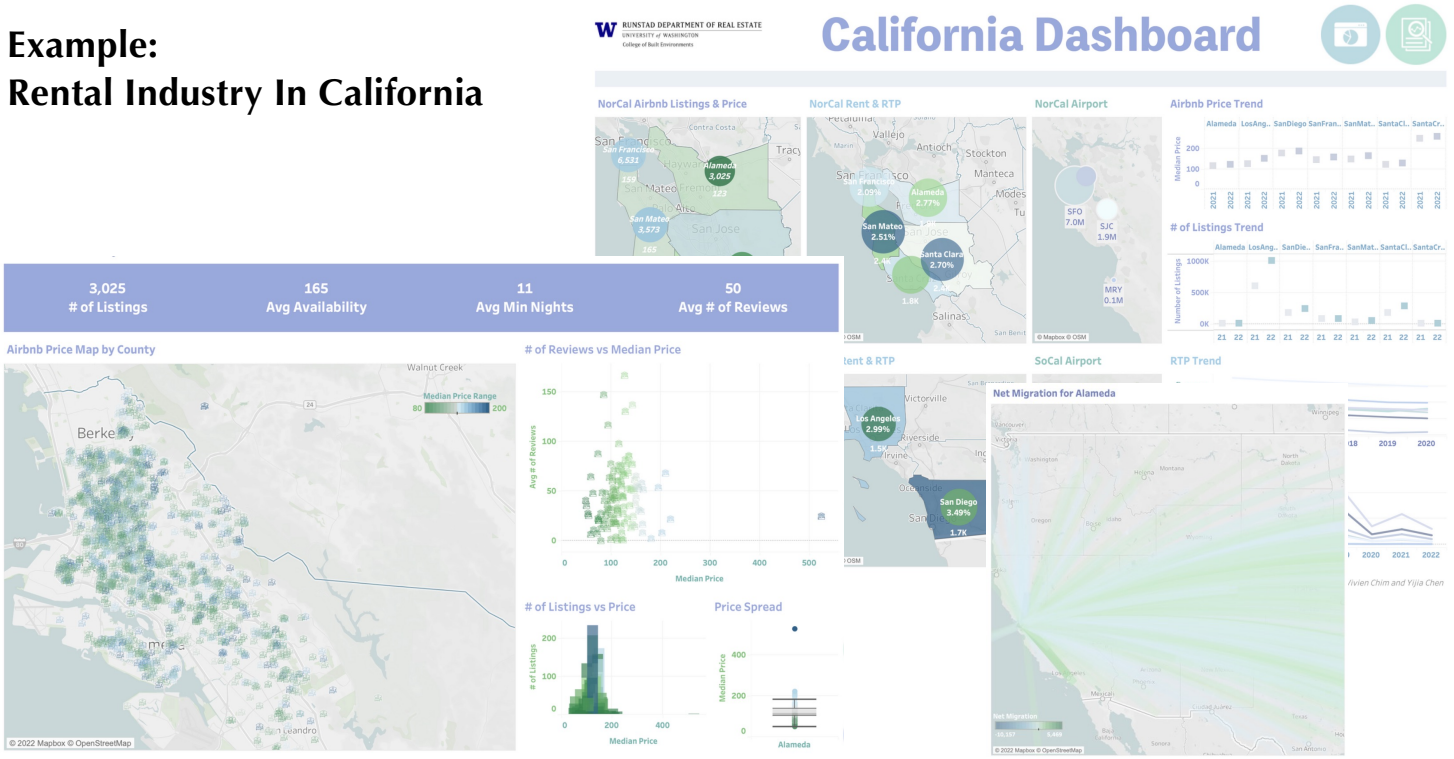


Data Analytics and Visualization Projects

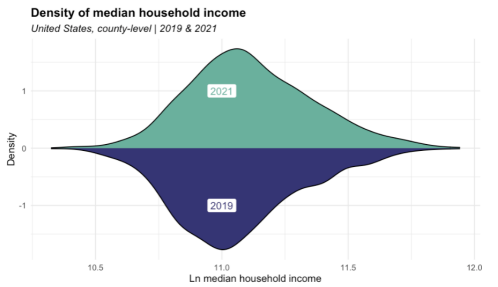
30% of the Total Grade (Website Link)

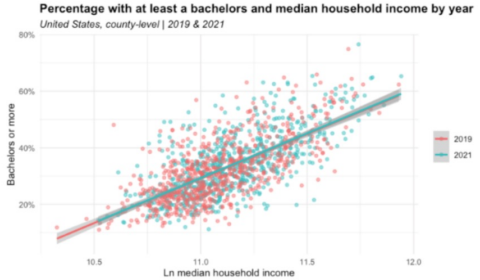
- 1~3 students for each group, and we expect everyone to spend 20~30 hours on the project.
- The project could be, for example:
  - A modeling of interesting datasets to derive new insights
  - Pure visualization for some datasets
  - A replication of an interesting academic article
- The requirements and their dates:
  - **10/08 -** Team Formation (1%)
  - **11/12 -** Project proposal (1~2 pages; 5%), you can submit anytime earlier for feedback
  - **12/03 -** Draft work presentation (in the last class, graded by peers and instructors; 8%)
  - **12/12 -** Final delivery (could be any format, like report, website, poster; 15%)
  - **12/12 -** Peer Review (1% of the total grade)

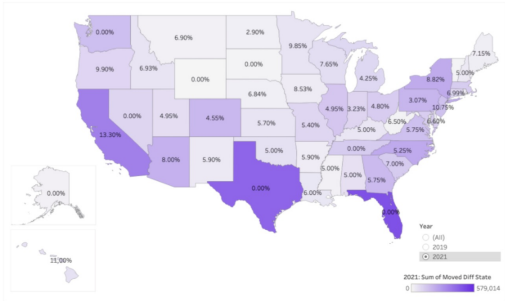
Example:
Rental Industry In California

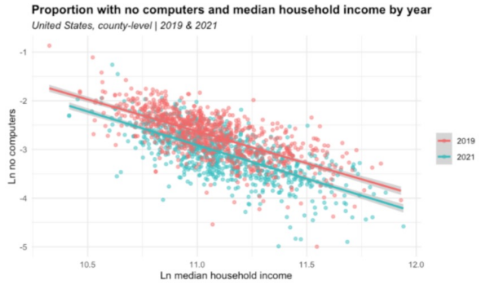


Example:
Tax-induced migration in the United States









Moved states | 2019 & 2021

	Dependent variable:			
	perMovedDiffState			
	w/o state (1)	w/o state (2)	w/ state (3)	w/ state (4)
NoIncomeTax	0.006***	0.007***	0.020**	0.027**
	(0.001)	(0.002)	(0.009)	(0.012)
HighestRate				-0.111
				(0.109)
year2021	0.002	0.002	0.002*	-0.003
	(0.001)	(0.001)	(0.001)	(0.002)
log(MedianHomePrice)				-0.0001
				(0.002)
log(MedianHHIncome)				-0.034***
				(0.004)
perBachelorsOrMore				0.054***
				(0.008)
perStudents				0.001
				(0.015)
perNoComputers				-0.201***
				(0.035)
NoIncomeTax:year2021		-0.001	-0.001	-0.001
		(0.002)	(0.002)	(0.003)
HighestRate:year2021				0.019
				(0.024)
Constant	0.026***	0.025***	0.027***	0.401***
	(0.001)	(0.001)	(0.002)	(0.043)
Observations	1,264	1,264	1,264	1,114
Adjusted R2	0.018	0.018	0.301	0.407
F Statistic	12.893***	8.684***	11.669***	14.375***
Note:			*p<0.1; **p<0.05; ***p<0.01	

Reminders

- Start to think about the final project and form groups (Oct 8, this Wednesday).
- Lab 1 will be due TODAY (Oct 6).
- CoStar access: fill out the form by TODAY!

Thank you!

Haoyu Yue / yohaoyu@washington.edu
Ph.D. Student, Interdisciplinary Urban Design and Planning
University of Washington

RE 519 Real Estate Data Analysis and Visualization
Course Website: www.yuehaoyu.com/data-analytics-visualization/
Autumn 2025

The course was developed based on previous instructors: Christian Phillips, Siman Ning, Feiyang Sun
Cover page credits: Visax