

4. Workflows and Data Visualization

The purpose of computing is insight, not numbers.

Richard Hamming

Purposes of Data Analysis in Urban Planning

Questions that data can answer, at least in part, differ by purpose. A **descriptive** question asks what happened. A **predictive** question asks what will happen. A **diagnostic** question asks why it happened: an associational analysis finds associations, and a causal analysis asks whether one thing caused another. A **prescriptive** question asks what we should do.

Descriptive: What Happened?

A descriptive analysis summarizes data that already exist. It does not intend to explain why a pattern exists, or to predict.

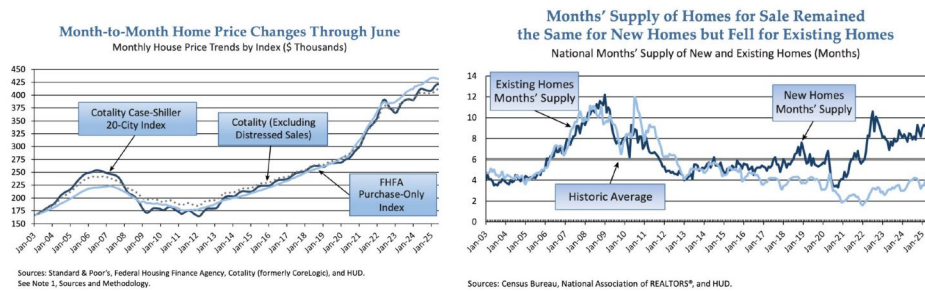


Figure 4.1: Housing market indicators, 2003-2025. Source: HUD Office of Policy Development and Research (2025), [Housing Market Indicators Monthly Update](#), August 2025.

With many variables and our limited mental capacity, description itself can also be based on models. For example, the study below uses two such methods to identify patterns within housing markets in Germany. **Principal component analysis (PCA)** combines correlated variables into a few components; here, 17 market variables become five. **Clustering** then groups places with similar values; here, German housing markets fall into four groups.

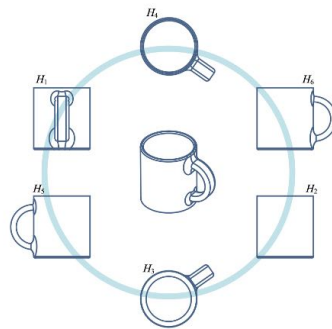
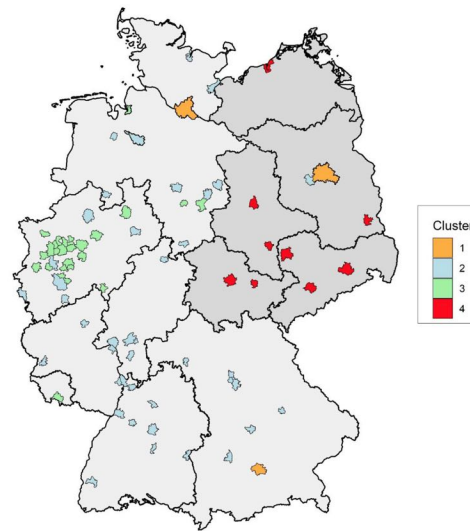


Figure 4.2: PCA illustration: a mug seen from six directions (H1-H6) around a circle, showing how different projections of the same object reveal different information. Source: [Visualize ML](#).

Example 4.1. **OPTIONAL** Grouping German housing markets using 17 Variables → Principal Component Analysis (PCA) → Reduced to 5 PCs → K-means Clus-

tering → **Example Conclusion:** German residential markets can best be segmented into four groups (Wiersma et al. 2022).

- *Market size and dynamics*
 - Population; Population growth (%); Housing stock (no. of flats); Vacancy rate (%)
- *Price level and dynamics*
 - Rents (EUR/sqm); Rental growth (%); Purchase prices condominiums (EUR/sqm); Purchase price growth (%); Gross initial yield (%)
- *Socioeconomic indicators and dynamics*
 - Purchasing power per household (EUR); Purchasing power growth (%); Unemployment (%); Rent affordability (%); Price affordability; Ownership rate (%)
- *Demographics*
 - Age cohort: 18-35 years (%); Change of age cohort: 18-35 years (%)



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

Figure 4.3: Housing market segments in Germany: regions colored by K-means cluster (1-4).

Predictive: What Will Happen?

A predictive analysis uses past data to estimate values not yet observed, such as the future sale price of a house. Prediction models, often machine learning algorithms, are judged mainly by their **prediction accuracy** on data they have not seen, not by whether the results can let us understand the relationship.

Example 4.2. OPTIONAL The study (Ho et al. 2021) utilizes three machine learning techniques to predict the transaction price of the property in Hong Kong, using floor area, age of the property from occupation permit to sale, floor level, facing direction, and travel time to the CBD.

The models predict prices well, but they do not tell us how each variable is related to the price. **Example Conclusion:** Each method can achieve an accuracy at XX levels in property price prediction.

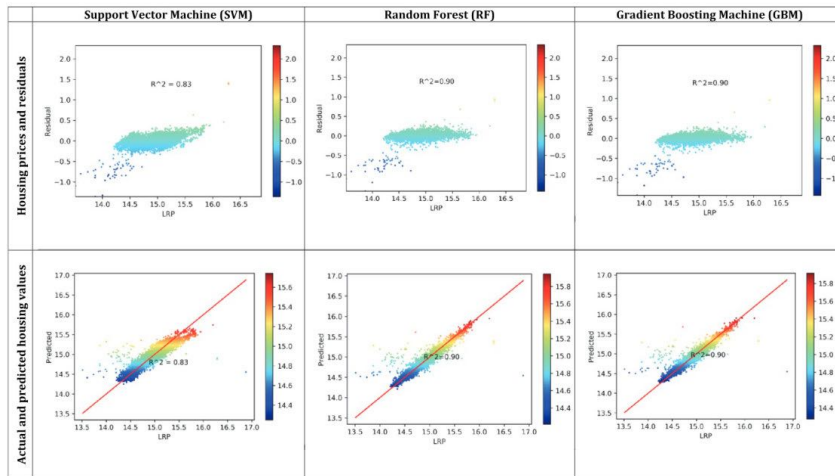


Figure 4.4: Actual vs predicted housing prices for three models: ($R^2 = 0.83$), ($R^2 = 0.90$), and gradient boosting machine ($R^2 = 0.90$).

Diagnostic (Associational): Why Did It Happen?

An associational analysis estimates how an outcome moves with other variables, holding the rest of the model fixed. It finds associations, so it answers “why” only in part. Across countries, for example, chocolate consumption per person is strongly correlated with the number of Nobel laureates per 10 million people (Figure 4.5). But we cannot say that in order to win more Nobel prizes, governments should start implementing chocolate promotion campaigns.

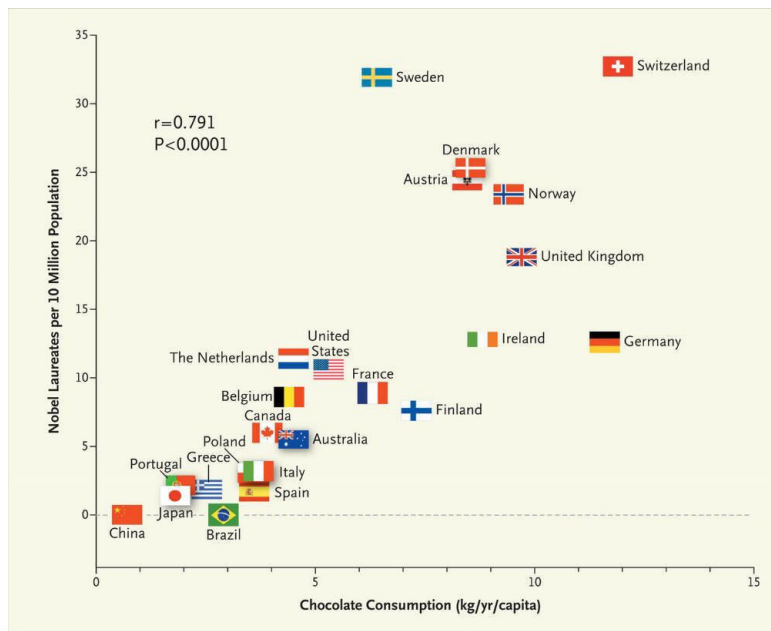


Figure 4.5: Chocolate consumption per person and Nobel laureates per 10 million people, by country. Source: Messerli (2012), redrawn by the [data.europa.eu data visualisation guide](https://data.europa.eu/data-visualisation-guide).

Example 4.3. OPTIONAL This study (Thompson et al. 2023) asks how flood events and flood risk are associated with house prices, comparing two sales of the same property. **Example Conclusion:** the statistically significant relationship (**associations**) between housing prices and other variables.

Hedonic price model coefficients	Coefficients	Significance
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 $\times$ Risk High	-6.6%	0.051
α_1 Bracket $\times$ Flood Zone	-9.4%	0.092
α_2 Bracket $\times$ Flood Zone $\times$ Risk High	-21.8%	< 0.001
α_3 Bracket $\times$ Flood Zone $\times$ Flood History	19.0%	0.034
α_4 Bracket $\times$ Flood Zone $\times$ Flood History	-12.5%	< 0.001
α_5 Bracket $\times$ Flood Zone $\times$ House Type	9.1%	0.215
α_6 Bracket $\times$ Flood Zone $\times$ 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

Diagnostic (Causal): Why Did It Happen?

A causal analysis asks whether an intervention changed the outcome. The ideal design is a controlled experiment, like a clinical trial, in which units are randomly assigned to receive the intervention or not. In most social science and planning questions, such experiments are impossible, too costly, or unethical, so we rely on observational data and designs that approximate an experiment.

Difference-in-differences (DID) is a common one. It compares the change in a group exposed to the intervention with the change in a similar group that was not; the second group's change stands in for what would have happened without the intervention.

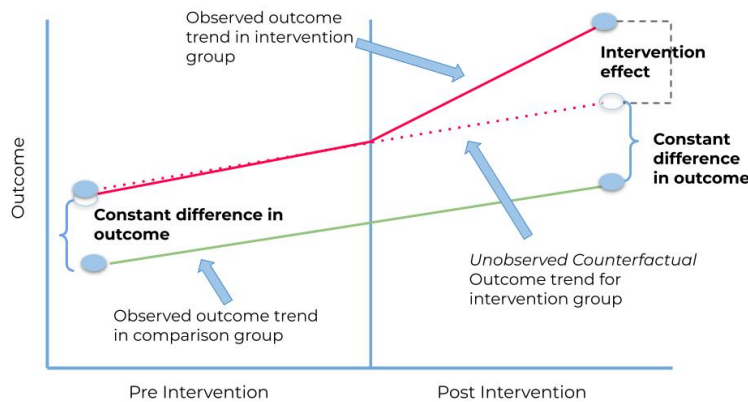


Figure 4.6: Difference-in-differences: the comparison group's trend stands in for the unobserved counterfactual of the intervention group, and the gap after the intervention is the effect. Source: [Bukalapak Data, Medium](#).

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 (Cao et al. 2015).

Prescriptive: What Should We Do?

A prescriptive analysis compares possible actions and recommends one. It often simulates many future scenarios and reports the spread of outcomes for each option, so a decision can weigh expected return against risk. Leung (2014) compares five designs for one office tower by their simulated net present value (Figure 4.7).

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

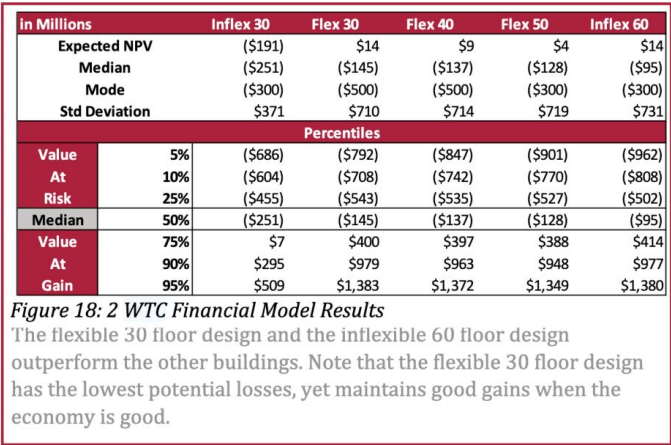


Figure 4.7: 2 WTC financial model results (in millions): expected Net Present Value (NPV) and value-at-risk / value-at-gain for flexible and inflexible building designs.

General Workflow of Working with Data

The questions can be answered through a common, but not fixed, sequence of stages, from a question about the real world to a result someone can use. The sequence is also not a straight line but an iterative process.

- **Define Questions** – Real World → Black Box
- **Data Collection and Management**
- **Data Cleaning** – Raw Data → Processed Data
- **Data Exploration** – Exploratory Data Analysis
- **Data Transformation** – Transformation; Feature Extraction
- **Data Modeling** – Descriptive; Predictive; Associational / Causal; Prescriptive
- **Interpretation and Communication** – Visualization and Reporting

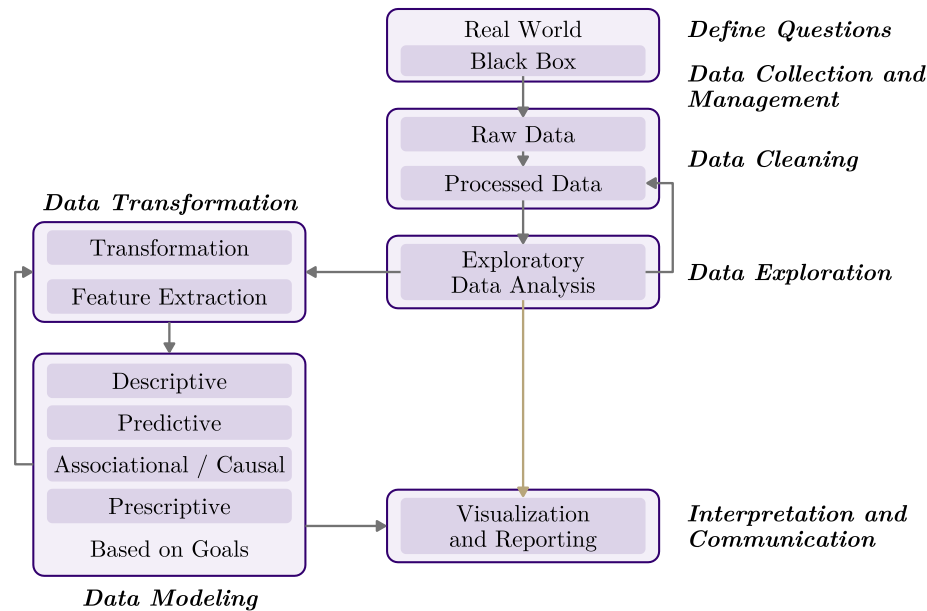


Figure 4.8: General workflow: overview.

From Real World to Data Representation

The first four stages turn a question into a dataset we understand. The real world is a black box: we never see it directly, only through the data we choose to collect. They usually take most of a project's time, and the choices made here shape every result that follows.

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
- Collect data and IRB of human-subjects involved

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

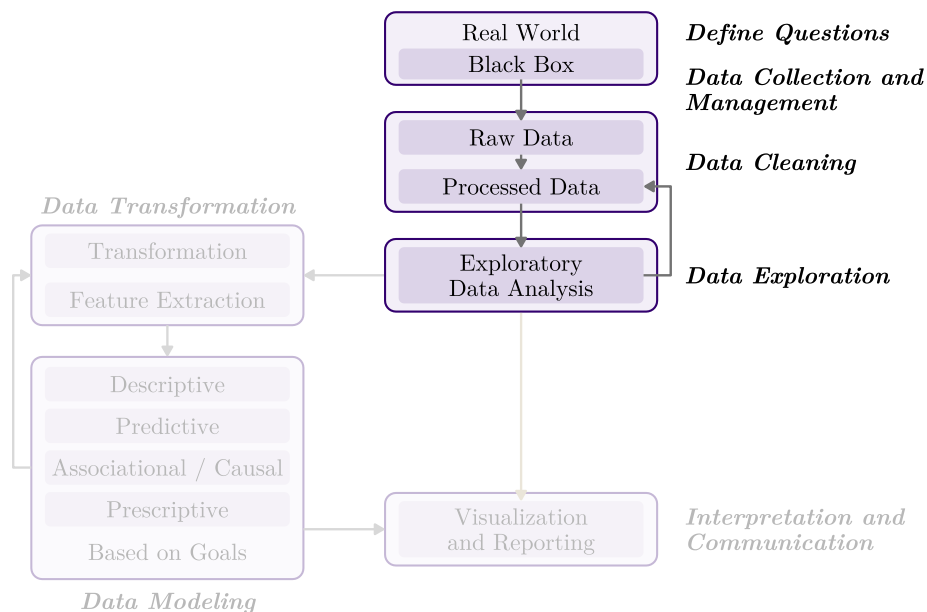


Figure 4.9: Define Questions, Data Collection and Management, Data Cleaning, and Data Exploration.

Data Modeling

Transformation prepares the variables, and the model is then chosen by purpose.

Data transformation before modeling:

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

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

- **Descriptive** (summarize what the data show)
 - If the data cover the whole population, the description is the answer. If they are a sample, add inference about the population.
- **Predictive** (estimate values not yet observed)
 - Judged by accuracy on data the model has not seen, not by whether its coefficients can be interpreted.
- **Associational** (how the outcome moves with other variables)
 - Association, however strong or significant, is not causal relationship.
- **Causal** (what an intervention changes)
- **Prescriptive** (choose the best option)
 - Usually builds on a predictive or causal model, since choosing the best option requires knowing what each option would do.

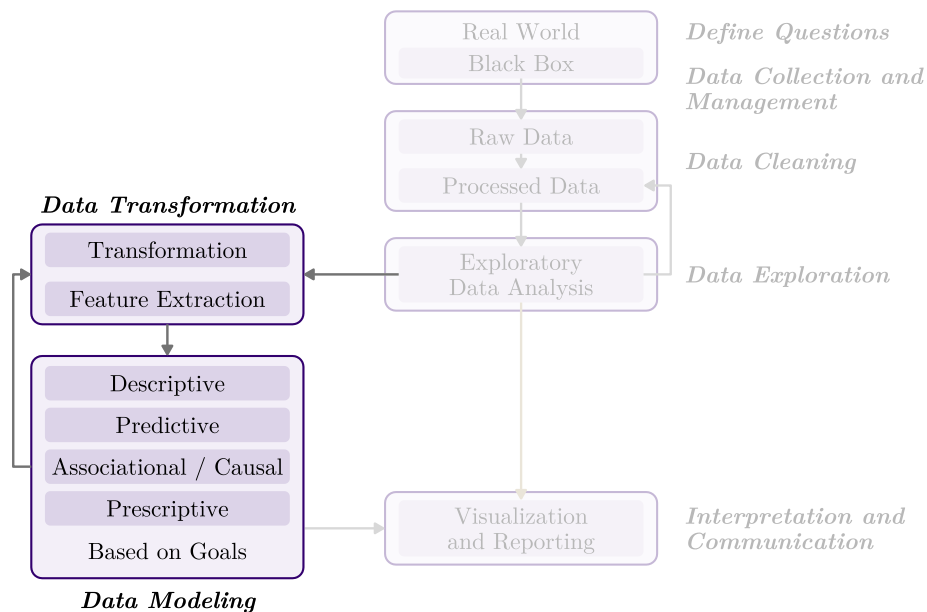


Figure 4.10: Data Transformation and Data Modeling.

Visualization and Communication

An analysis is finished only when someone else can understand and use it. The last stage turns results into a message for a specific audience.

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

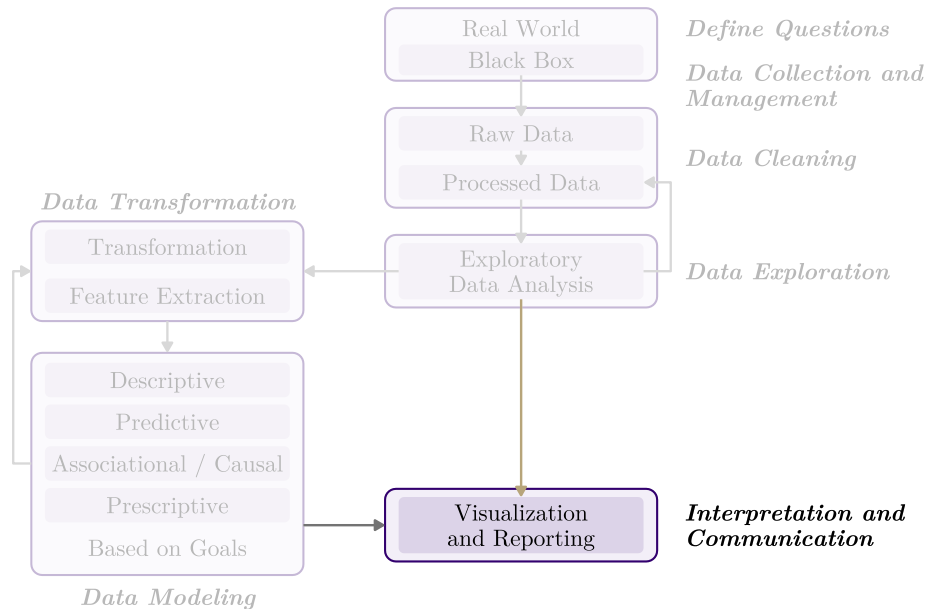


Figure 4.11: Visualization, Interpretation and Communication.

Data Visualization

A good figure can show a pattern that a table hides. Visualization serves two jobs in the workflow: while exploring, it helps us find what we did not expect; when communicating, it carries a result to an audience.

Cholera Map

Cholera broke out in the Broad Street area of central London on the evening of August 31, 1854. John Snow, who had investigated earlier epidemics, suspected that the water from a community pumpwell at Broad and Cambridge Streets was contaminated. Testing the water from the well on the evening of September 3, Snow saw no suspicious impurities, and thus he hesitated to come to a conclusion ... (Tufte 1997)

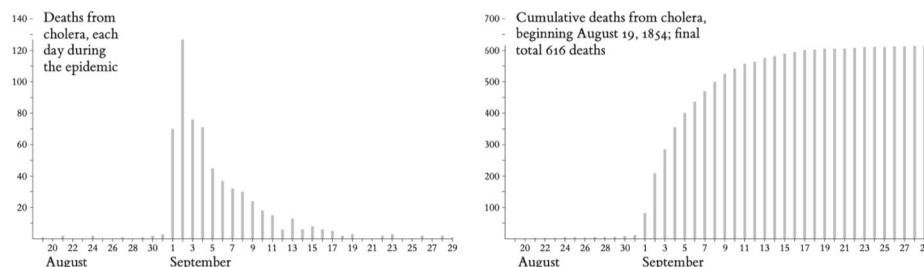


Figure 4.12: Deaths from cholera each day and cumulative deaths, London, August 20 to September 29, 1854; 616 deaths in total. Source: Tufte (1997).

Snow then mapped the deaths, marking each one as a black bar at the victim’s address (Figure 4.13). The deaths cluster around the Broad Street pump, which pointed to the pump as the source of the outbreak.



Figure 4.13: John Snow’s 1854 map of the Broad Street area of London; each black bar is one cholera death. Source: Tufte (1997).

Explorative vs. Narrative

Can you tell what the differences are between these two figures?

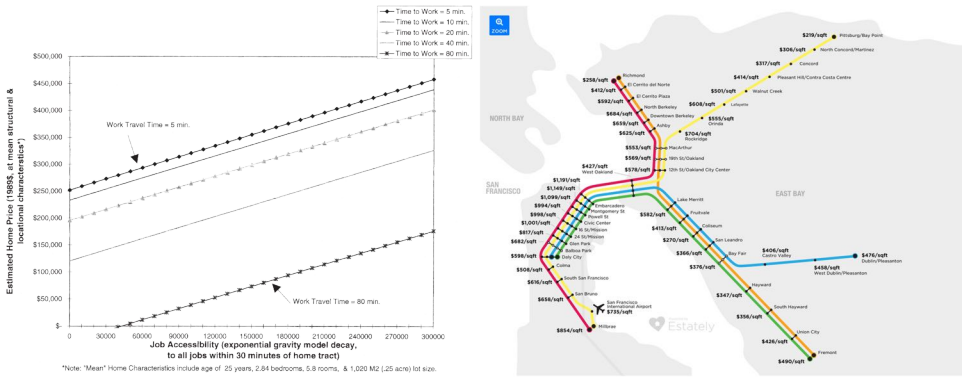


Figure 4.14: Left: house prices versus job accessibility at different work travel times. Source: Kockelman (1997). Right: San Francisco Bay Area housing prices per square foot on a BART map. Source: Reddit, original post by /u/surf2japan in r/bayarea.

The left figure is **explorative**: it encodes only data and is written for researchers who will read it closely. The right figure is **narrative**: it borrows the BART map to catch a general reader’s eye, at the cost of precise comparison. Lankow et al. (2012) distinguish the two kinds by purpose:

	Explorative	Narrative
Characteristics	Minimalist Only includes elements that represent data	Illustrative Design-focused

	Explorative	Narrative
	Seeks to communicate information In the most clear, concise manner	Seeks to appeal to viewer with engaging visuals Informs and entertains
Applications	Academic research; science; business intelligence; data analysis	Publications; blogs; content marketing; sales and marketing materials

Types of Data Visualization

Static

A **static** visualization is a still image that readers view and read (Figure 4.15). It works best as a narrative, though it can serve exploration. Most reports, plans, and slide decks use static figures, as does editorial content for blogs and social media, and they go out of date as soon as the data change.

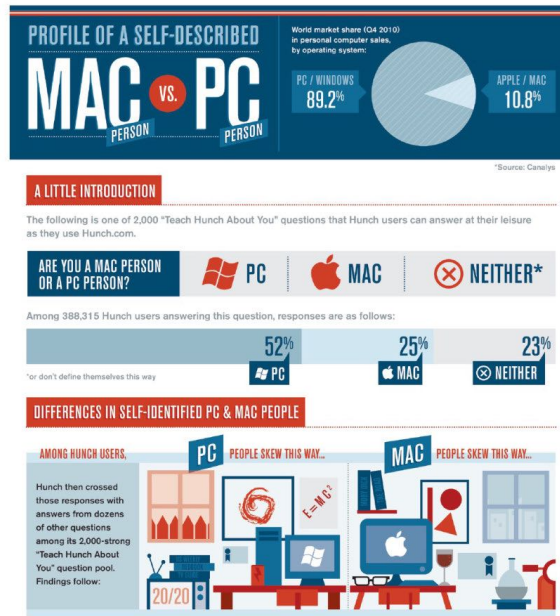


Figure 4.15: Profile of a self-described Mac vs. PC person, a static infographic. Source: Lankow et al. (2012, ch. 2, p. 67).

Motion

A **motion** visualization is animated, sometimes with a voiceover (Figure 4.16). It is almost always narrative: the viewer cannot stop to explore unless the animation is also interactive.

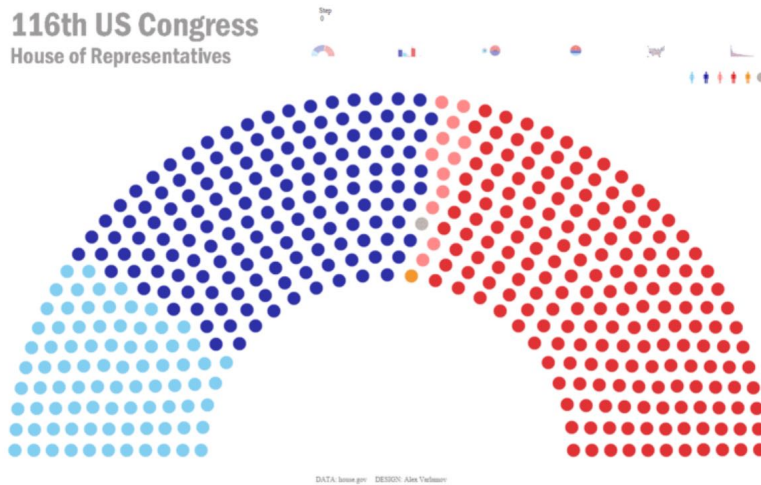


Figure 4.16: A frame of an animated Tableau visualization of the 116th US House of Representatives, seats colored by party. Source: Flerlage Twins.

Interactive

An **interactive** visualization lets the reader click, filter, search, and choose what to show, so it can be explorative, narrative, or both (Figure 4.17). Public agencies use dashboards and online maps for internal reports and public engagement. An interactive tool that exposes records also raises questions of data security and privacy.

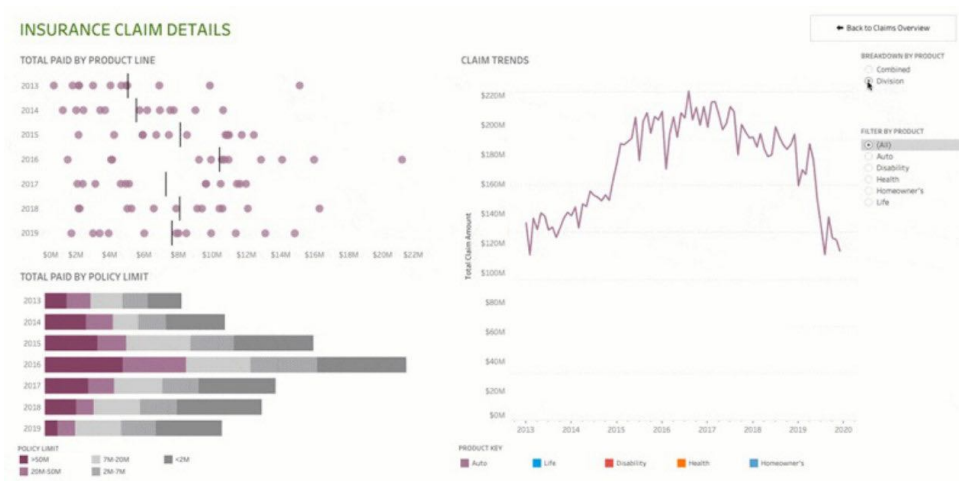


Figure 4.17: An interactive Tableau dashboard of insurance claim details, with filters by product and breakdown. Source: Tableau Help.

More Examples of Data Visualization

💡 Browse more chart types

We strongly recommend two online catalogues. The [Data Viz Project](#) collects chart types with examples, and lets you filter them by what a chart does.

From [Data to Viz](#) starts from your data instead: choose the kind of variables you have, and a decision tree suggests charts that fit, with common pitfalls and code in R and Python.

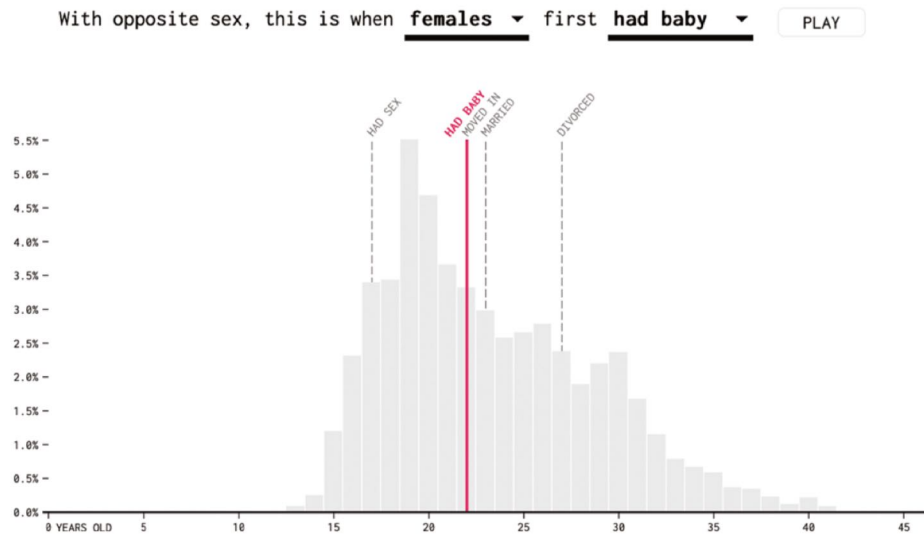


Figure 4.18: When females first had a baby: share of people by age, an interactive histogram. Source: Yau (2024).

JOB OF PHYSICIAN SPOUSES

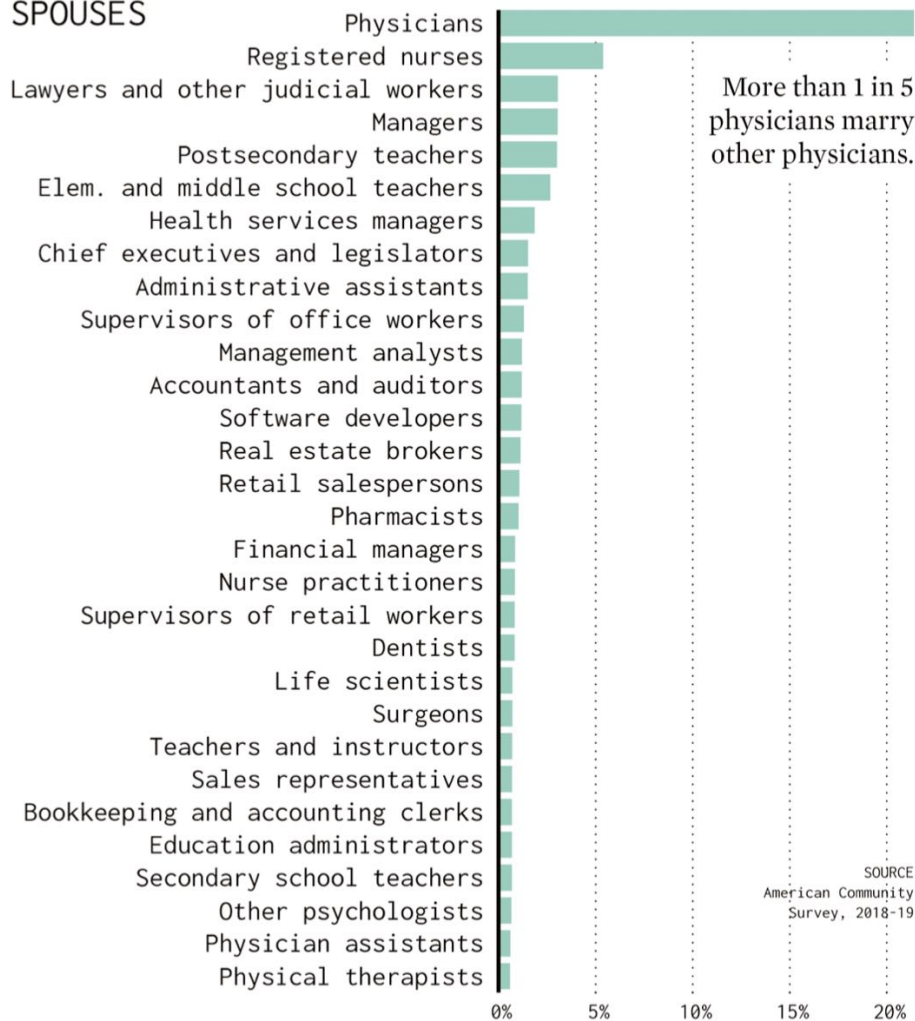


Figure 4.19: Jobs of physician spouses: more than 1 in 5 physicians marry other physicians (American Community Survey, 2018-19). Source: Yau (2024).

PEOPLE MARRYING LATER

Each line represents a decade from 1900 to 2021. Thicker lines are more recent.

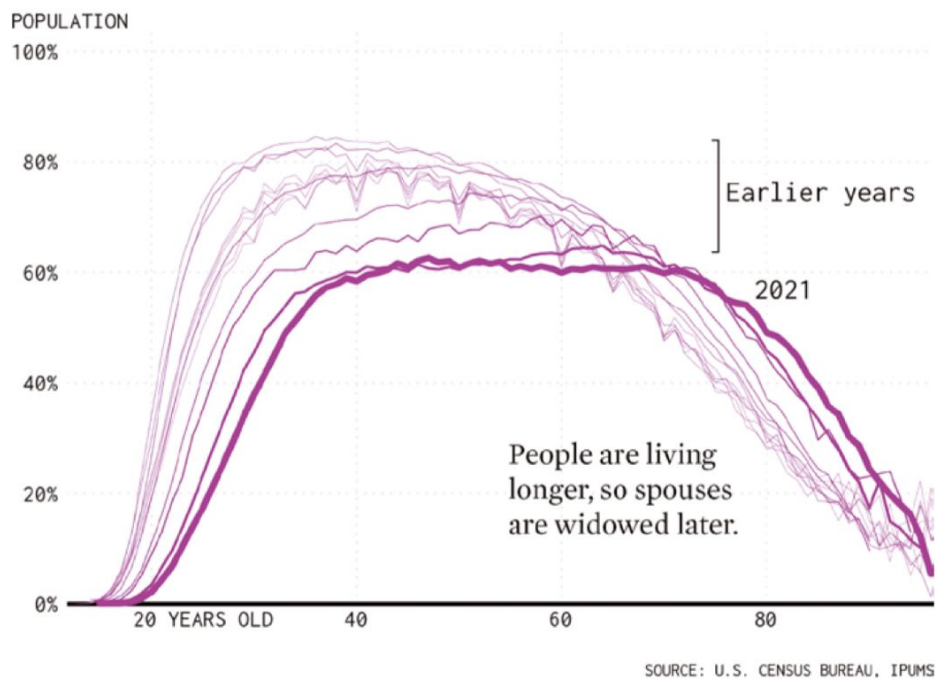


Figure 4.20: People marrying later: percentage married by age, one line per decade from 1900 to 2021 (U.S. Census Bureau, IPUMS). Source: Yau (2024).

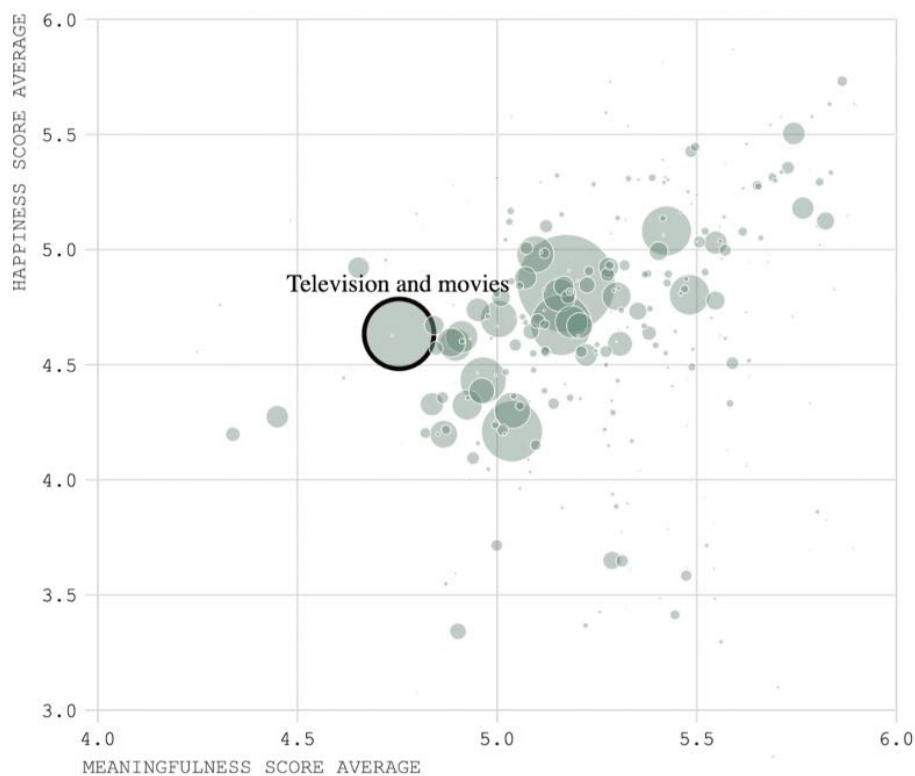


Figure 4.21: Meaningfulness and happiness of daily activities, a bubble chart. Source: Yau (2024).

The bluer the district, the more valuable the primary vote

The average blue congressional district awards one convention delegate per 28,912 Romney voters, while the average red district awards one delegate per every 56,714 Romney voters

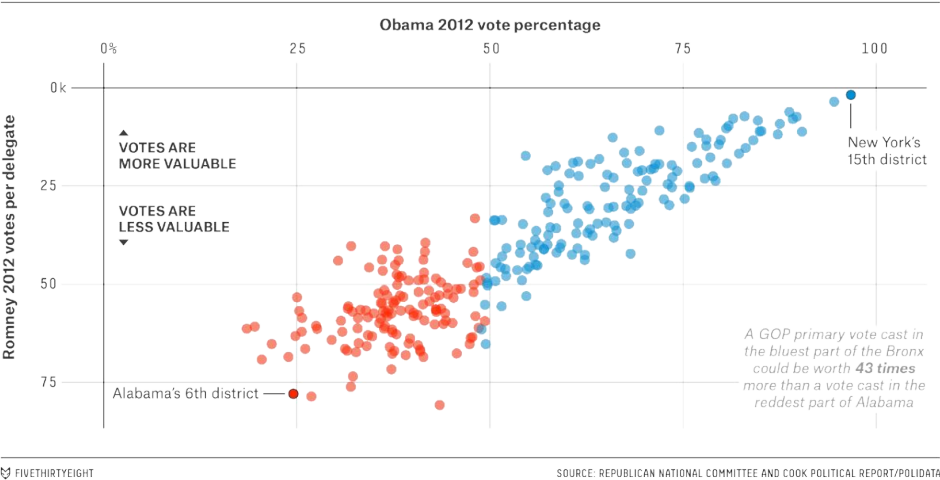


Figure 4.22: The bluer the district, the more valuable the primary vote. Source: FiveThirtyEight.

FATAL CRASHES INVOLVING ALCOHOL PEAK AT NIGHT

Based on 2015 data from the National Highway Traffic Safety Administration.

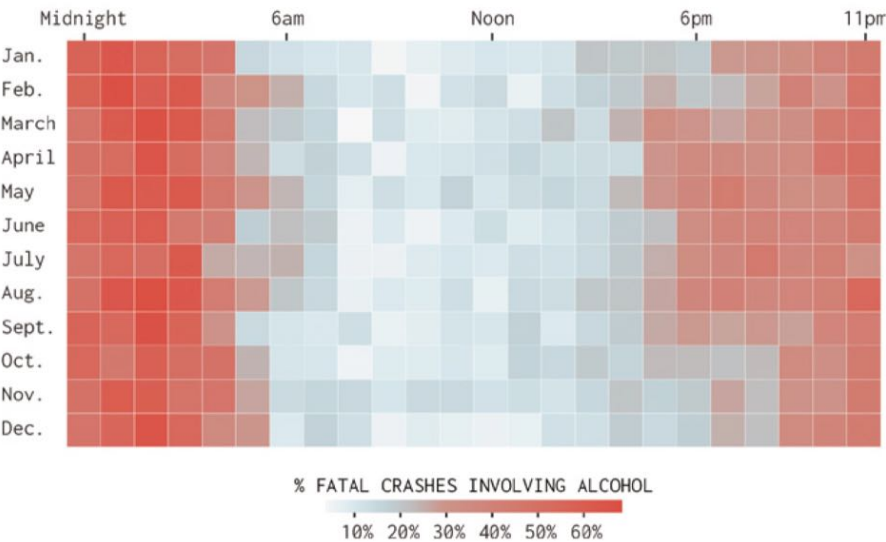


Figure 4.23: Fatal crashes involving alcohol by month and hour of day (2015 NHTSA data) Source: Yau (2024).

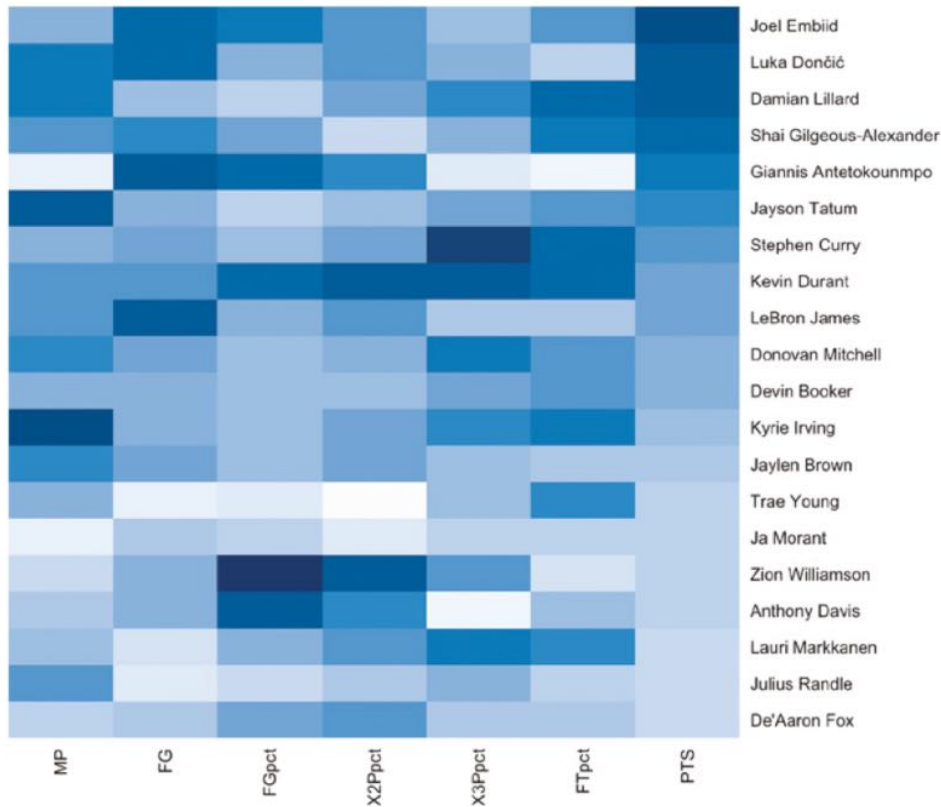


Figure 4.24: A correlation matrix of NBA player statistics. Source: Yau (2024).

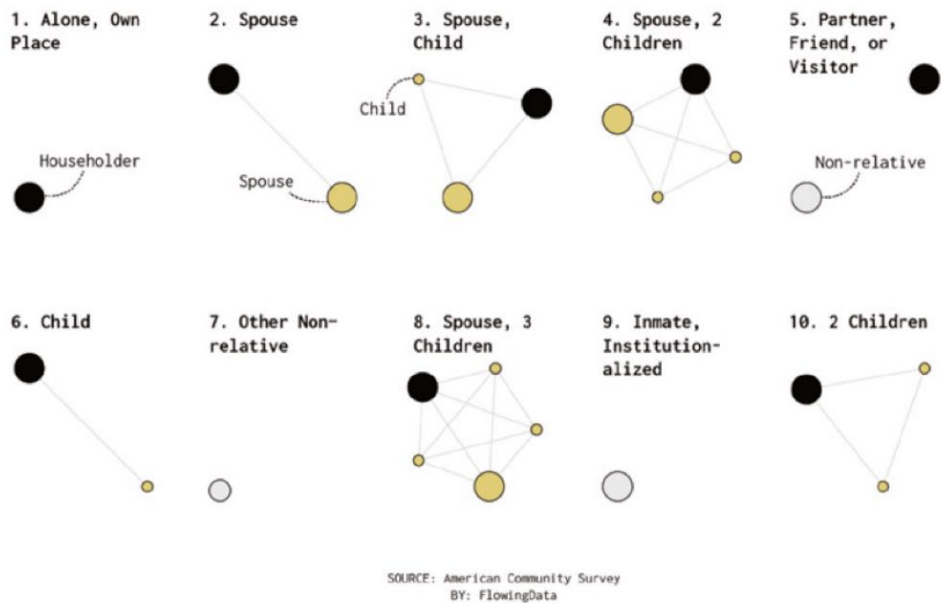


Figure 4.25: The ten most common household types drawn as small networks (American Community Survey). Source: Yau (2024).



Figure 4.26: Most common coffee chains in the United States, by the most popular chain within a 10-mile radius (AggData). Source: Yau (2024).

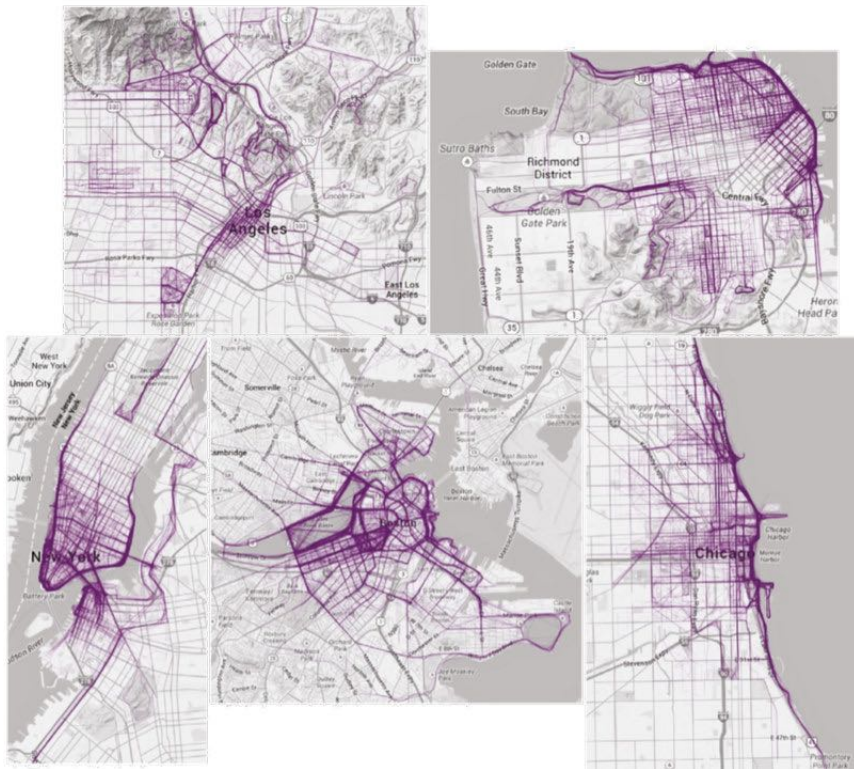


Figure 4.27: Where do people run in major cities? Popular running routes in Los Angeles, San Francisco, New York, and Chicago. Source: Nathan Yau, FlowingData.