

Module 2

Data Visualization

Shmueli et al., Chapters 2 and 3 (part)

Distributions

Relationships

Advanced & 3D

Principles

Roumen Vesselinov, PhD **Roumen@jhu.edu**

Explore the data · build the model · ask the AI · verify every result

Outcomes and predictors

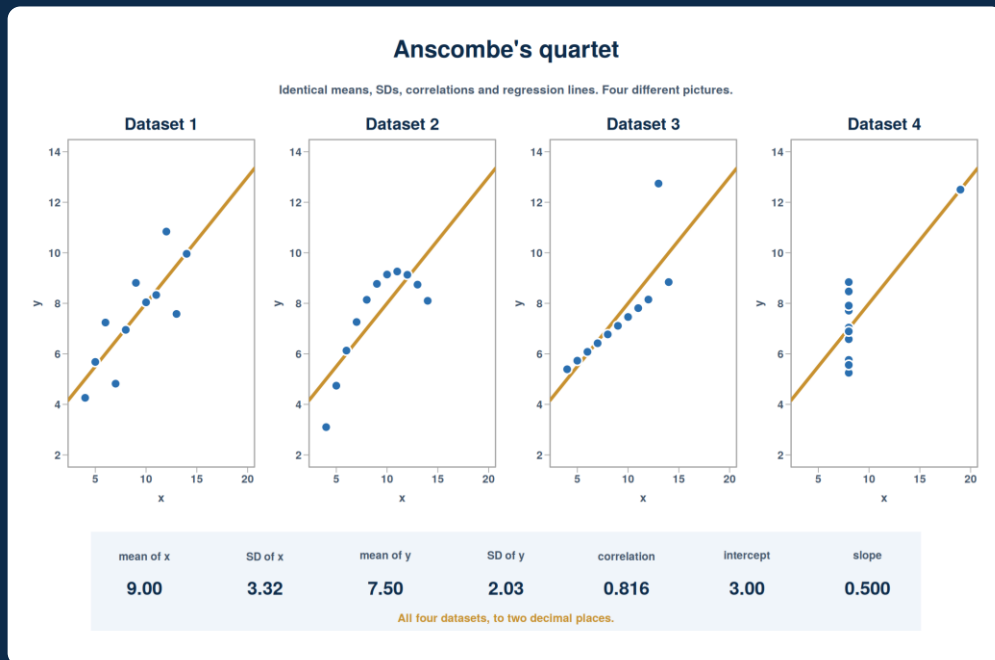
Var	Description	Var	Description
Y1	GDP per capita, constant 2015 US\$	Y2	Armed conflict (yes/no)
X1	Annual GDP growth rate (%)	X11	Total population
X2	Inflation rate (CPI, annual %)	X12	Population aged 65+ (% of total)
X3	Trade (% of GDP)	X13	Urban population (% of total)
X4	FDI net inflows (% of GDP)	X14	Top 10% income share (inequality)
X5	Fraser Economic Freedom Index	X15	Government effectiveness (WGI)
X6	Oil rents (% of GDP): resource channel	X16	Rule of law (WGI)
X7	Human Development Index	X17	Political stability & absence of violence (WGI)
X8	Life expectancy at birth (years)	X18	V-Dem electoral democracy index
X9	Internet users (% of population)	X19	Historical region (10 categories)
X10	Government expenditure on education (% of GDP)	X20	Year (numeric, for time-trend modeling)

X19: historical region

Value	Region	Value	Region
1	Eastern Europe & post-Soviet Union	6	East Asia
2	Latin America	7	South-East Asia
3	North Africa & Middle East	8	South Asia
4	Sub-Saharan Africa	9	The Pacific
5	Western Europe & North America (incl. Australia, NZ)	10	The Caribbean

X19 is categorical: in models it becomes dummy variables.

Summary statistics can hide everything



Anscombe's quartet (1973)

- Four data sets with the same means, SDs, correlation, and regression line.
- Plots reveal a curve, an outlier, and a single influential point.
- Always plot the data before modeling.

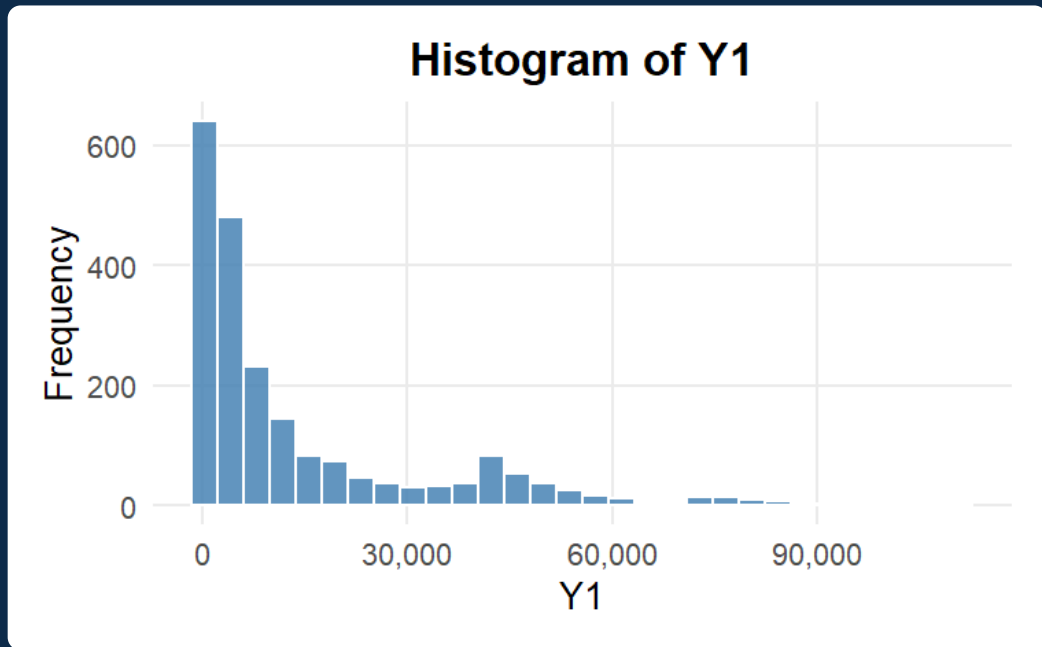
Which graph for which question?

Question	Graph	In R (ggplot2)
Distribution of one variable	Histogram, density, boxplot, violin, ECDF	<code>geom_histogram</code> , <code>geom_boxplot</code> , <code>geom_violin</code>
Compare groups	Side-by-side boxplots, violins, ridgelines	<code>aes(x = region, y = Y1); ggridges</code>
Relationship of two variables	Scatterplot with smoother, hexbin	<code>geom_point</code> + <code>geom_smooth</code>
Three or more variables	Bubble chart, 3D scatter, scatterplot matrix	<code>plotly</code> , <code>GGally::ggpairs</code>
Many correlations	Correlation heatmap	<code>corrplot</code> , <code>ggcorrplot</code>
Change over time	Line charts, small multiples	<code>geom_line</code> + <code>facet_wrap</code>

PART 1

Distributions of Y1: GDP per capita

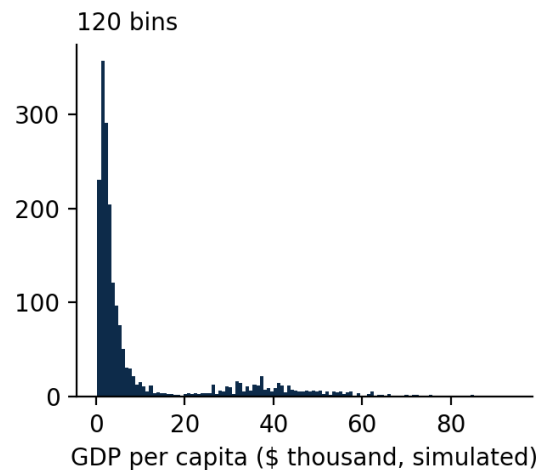
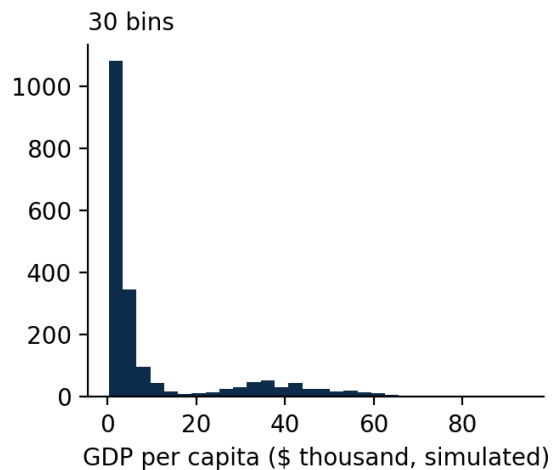
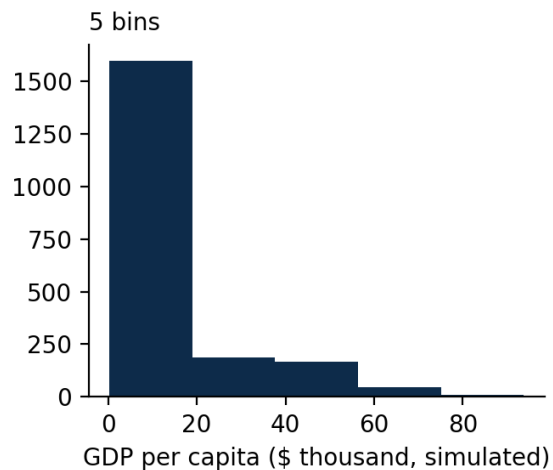
Histogram shows the distribution of the outcome variable



Histogram of Y1

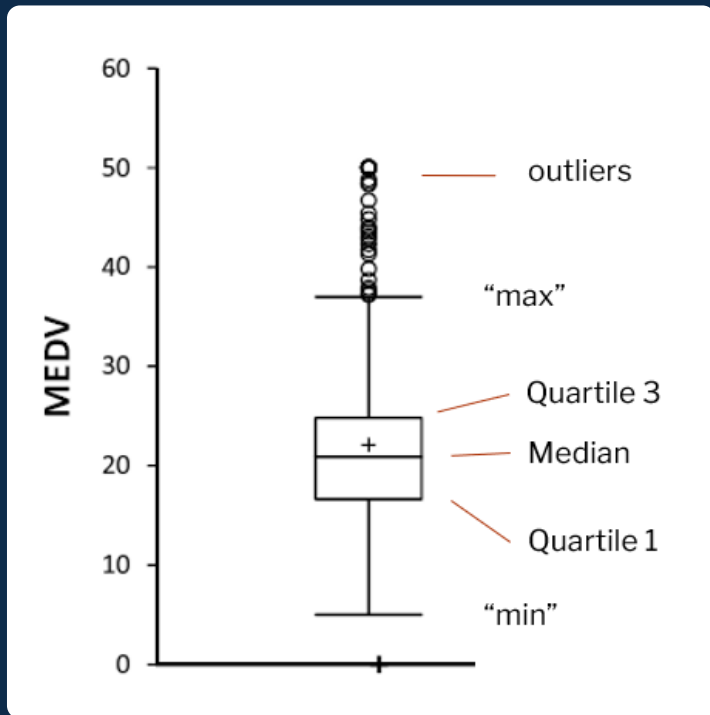
- Strong right skew: most country-years are poor, a few very rich.
- The mean lies above the median.
- A second, smaller group of rich countries appears around \$30,000–60,000.
- The shape depends on the bin width: try several.

Same data, different bins



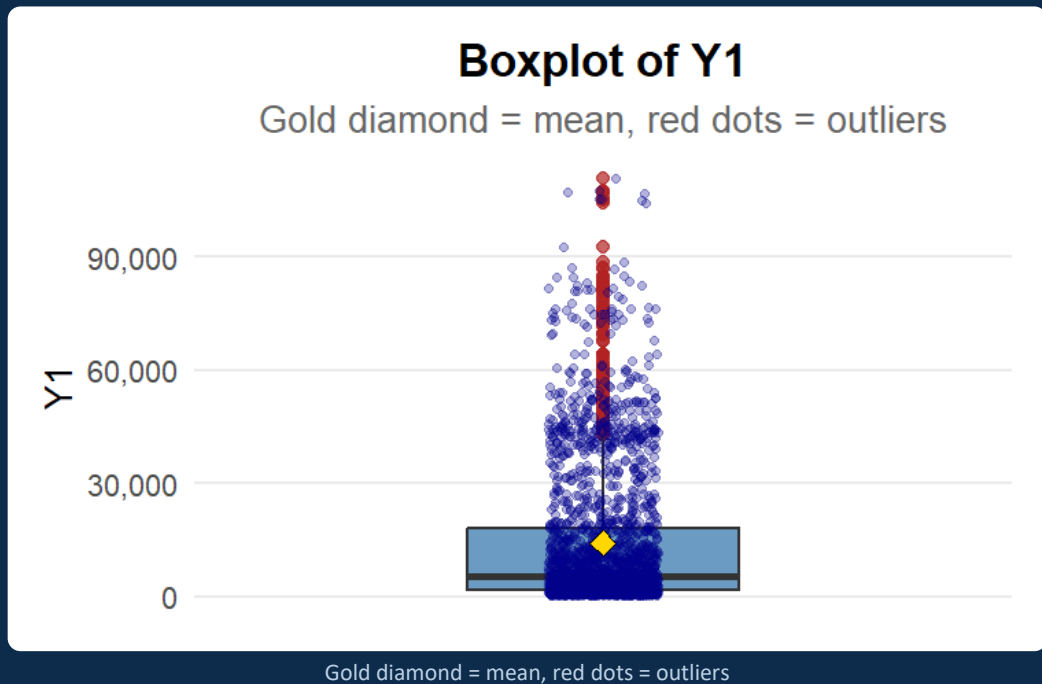
5, 30, and 120 bins (simulated GDP per capita)

Anatomy of a boxplot



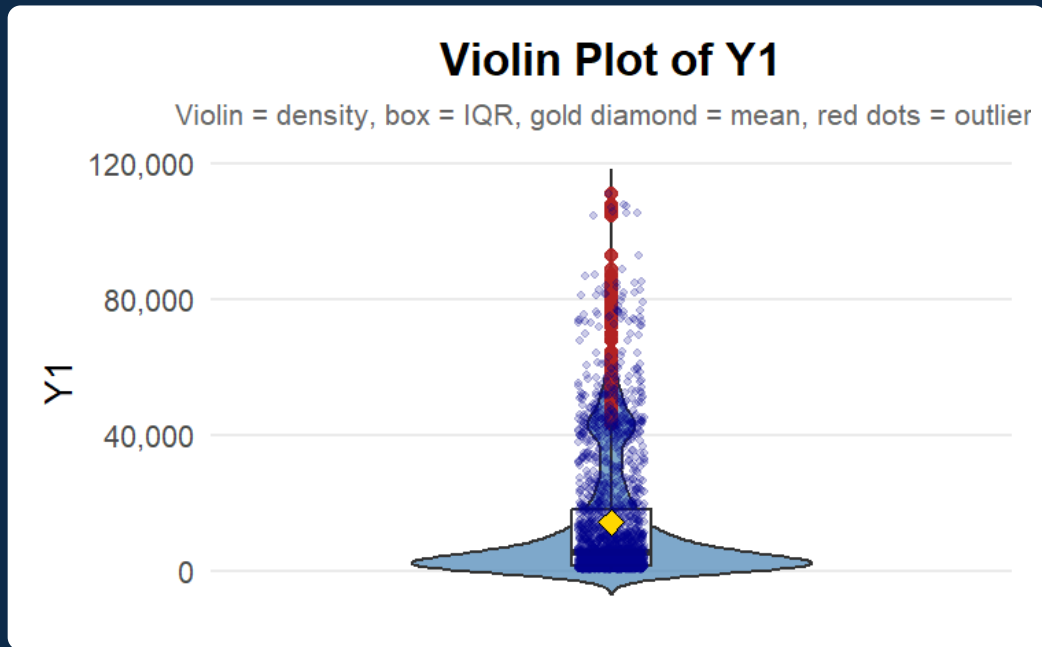
- Top outliers: values above $Q3 + 1.5(Q3 - Q1)$.
- "Max" = maximum of the non-outliers.
- Analogous definitions for bottom outliers and for "min".
- Box: quartile 1, median, quartile 3.
- Details may differ across software.

Boxplot of Y1



- The box is compressed at the bottom: half of country-years are below about \$5,000.
- Many points flagged as outliers: all rich countries.
- On skewed data, "outlier" $\neq$ error.

Violin plot of Y1



Violin = density, box = IQR, gold diamond = mean, red dots = outliers

- Wide base: most country-years have low GDP per capita.
- Long thin neck: a small group of rich countries.
- Mean (gold) well above the median: right skew.

How to read a violin plot

- A violin plot combines two views of a distribution in one figure.
- The shape (violin) is a mirrored density plot: wider regions mean more observations at that value, narrower regions fewer. Essentially a smoothed histogram turned sideways and reflected.
- The box inside is a standard boxplot: median line, box from the 25th to the 75th percentile (IQR), whiskers to the smallest and largest values within $1.5 \times \text{IQR}$ of the box.
- The diamond (gold) marks the mean.
- Red dots beyond the whiskers are outliers; jittered points show the raw observations behind the smoothed shape.

What you should look for

1

Widest part

The typical or most common values (the mode).

2

Asymmetry

Skewness: a long tail on one side means the distribution is skewed in that direction.

3

Multiple bulges

Multimodality: more than one cluster of values.

4

Mean vs. median

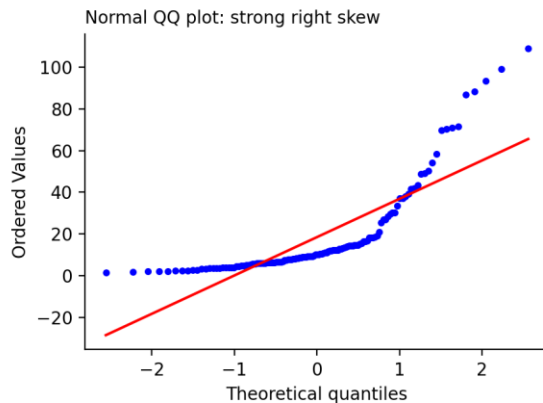
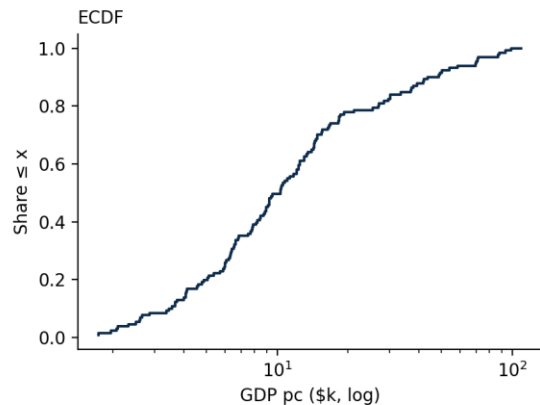
If the gold diamond sits noticeably above or below the median line, the distribution is skewed.

5

Outliers

Extreme values that may need investigation.

ECDF and QQ plot



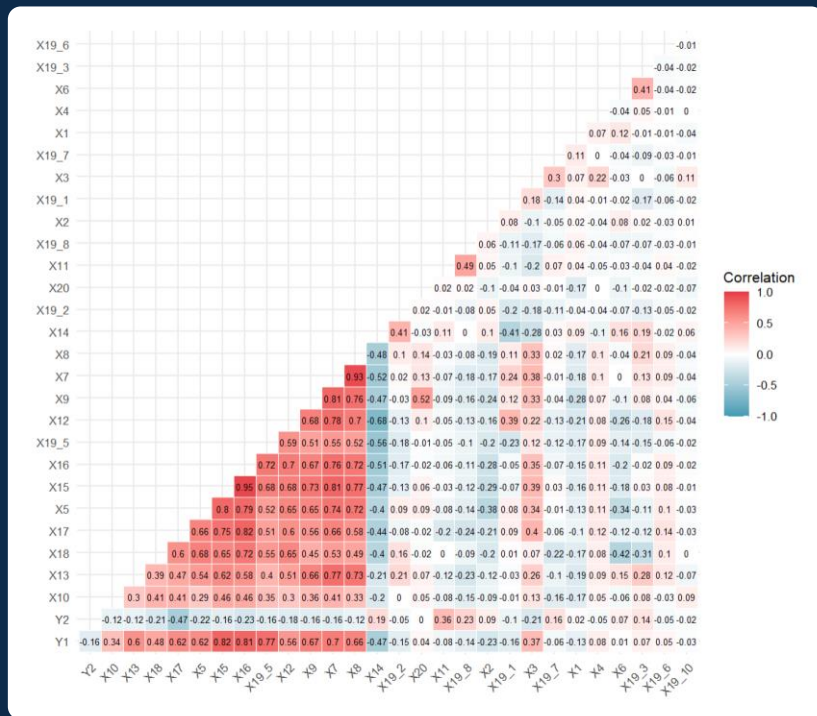
Simulated GDP per capita, 2021

- ECDF (empirical cumulative distribution function): share of observations $\leq$ each value; no bins needed.
- QQ plot: points bend away from the line $\rightarrow$ skew and heavy tails.
- QQ plots are also used to check regression residuals (Module 4).

PART 2

Relationships among variables

Correlation heatmap of GOV1 variables



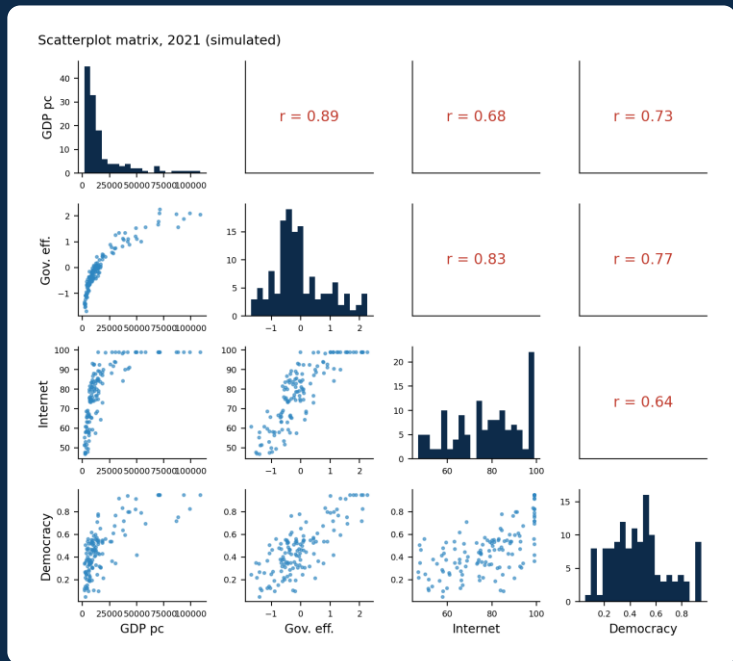
- Each cell: Pearson correlation of two variables.
- Red = positive, blue = negative; darker = stronger.
- Blocks of red among predictors warn of multicollinearity (Module 4).
- Only linear, pairwise; region (categorical) is not shown.

Top 6 correlates of Y1



- X15 government effectiveness (0.82) and X16 rule of law (0.81) are the strongest.
- They also correlate 0.95 with each other: nearly redundant.
- Follow up with scatterplots: strong correlation can hide curvature.

Scatterplot matrix



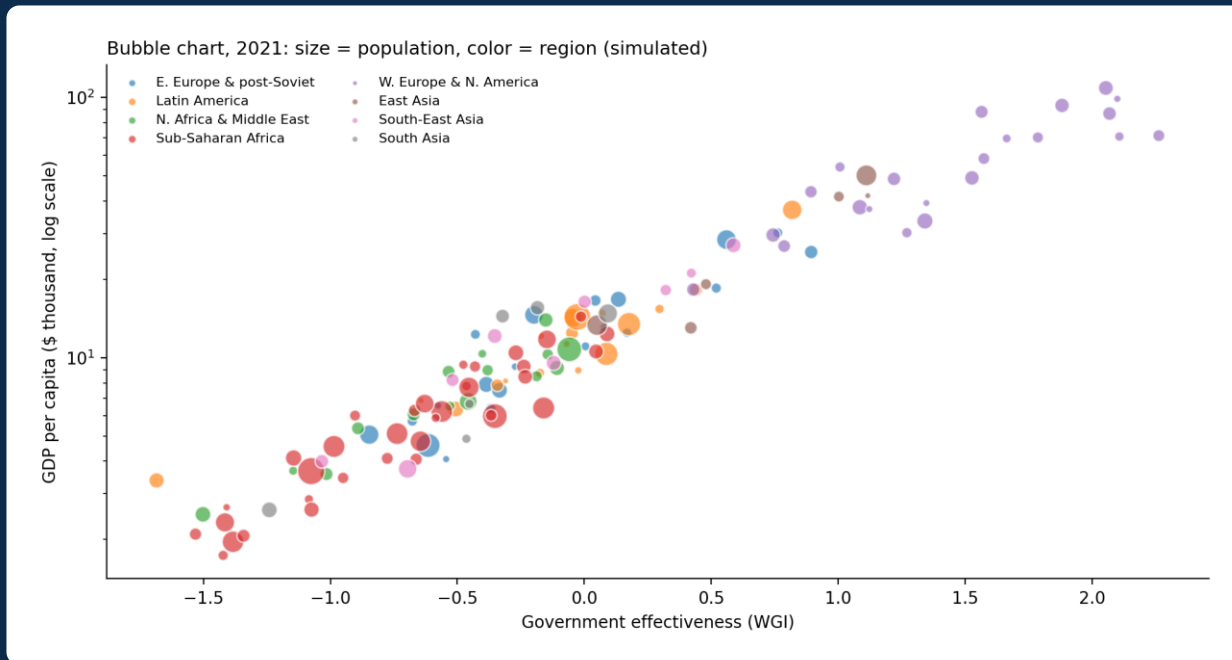
GGally::ggpairs (simulated data)

- Diagonal: each variable's distribution.
- Below: scatterplots; above: correlations.
- Reveals curvature and outliers a heatmap cannot show.
- Practical for up to about 8–10 variables.

PART 3

Advanced visualization tools

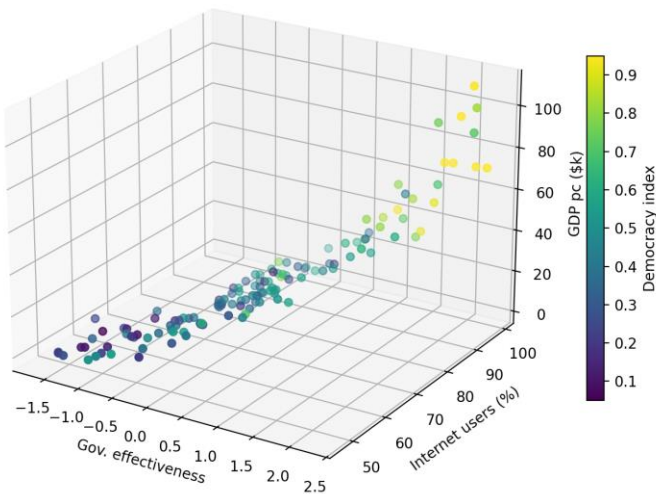
Bubble chart (Gapminder style)



x = government effectiveness, y = GDP per capita (log), size = population, color = region

3D scatterplot

3D scatterplot, 2021 (simulated)

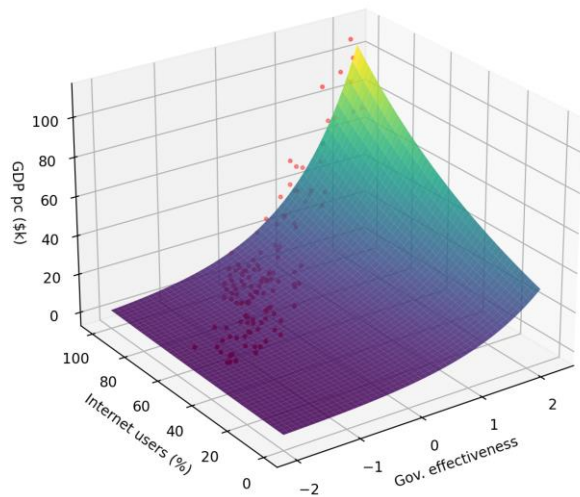


GDP per capita vs. government effectiveness and internet use; color = democracy (simulated)

- Shows how two predictors jointly relate to the outcome.
- Static 3D views can hide points: rotate interactively (plotly, rgl).
- Perspective distorts distances: use for exploration, not precise reading.

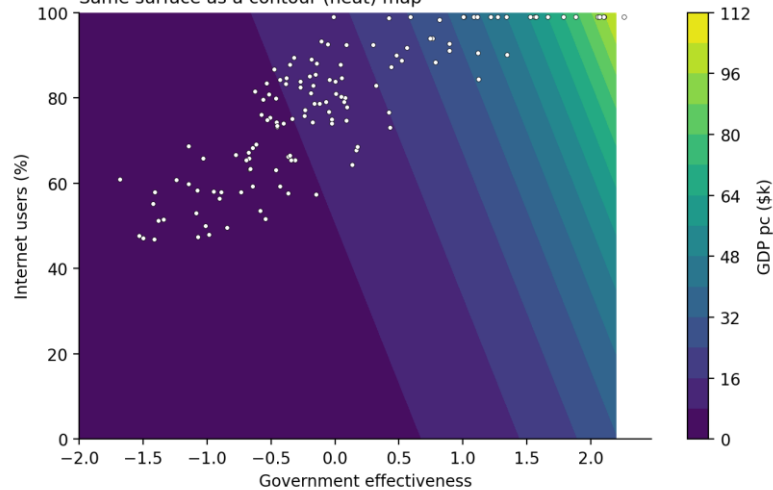
3D surface and contour map

3D surface: expected GDP per capita, with data points



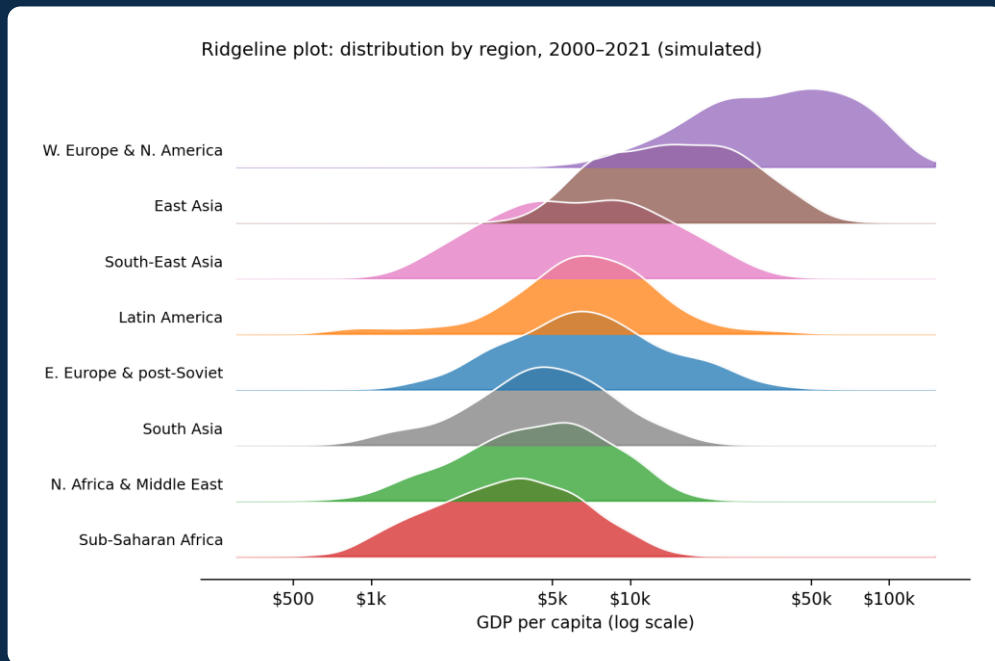
Surface: expected GDP per capita

Same surface as a contour (heat) map



Same surface as a contour map

Ridgeline plot

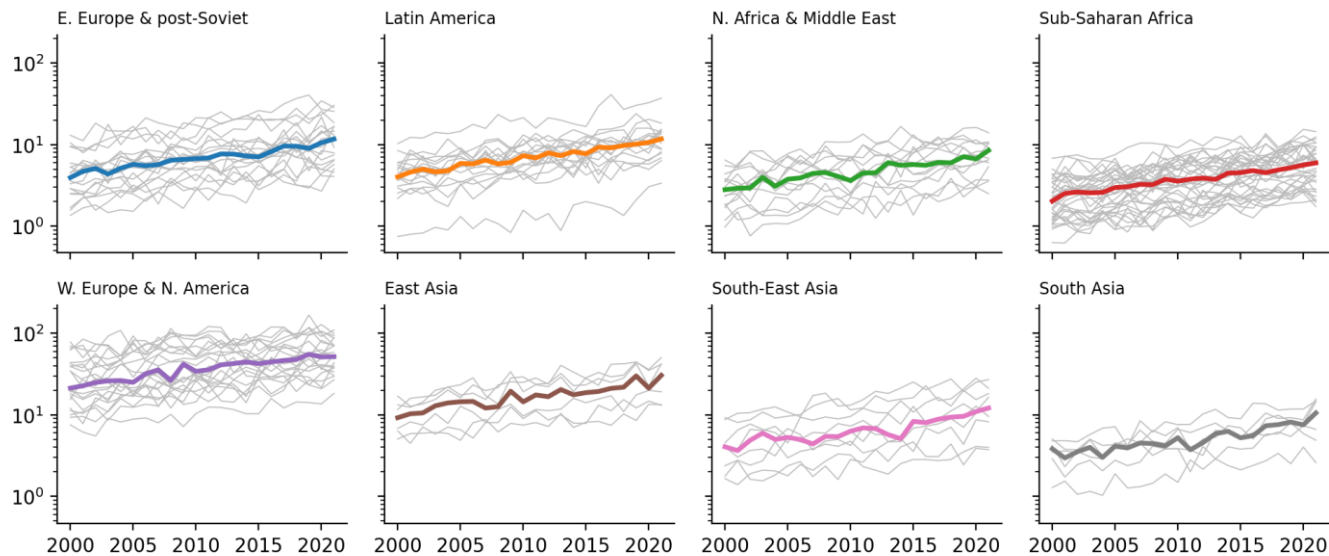


`ggridges::geom_density_ridges (simulated)`

- One density per region, stacked and ordered by median.
- Compares shapes, not just medians.
- Log scale makes skewed data readable.

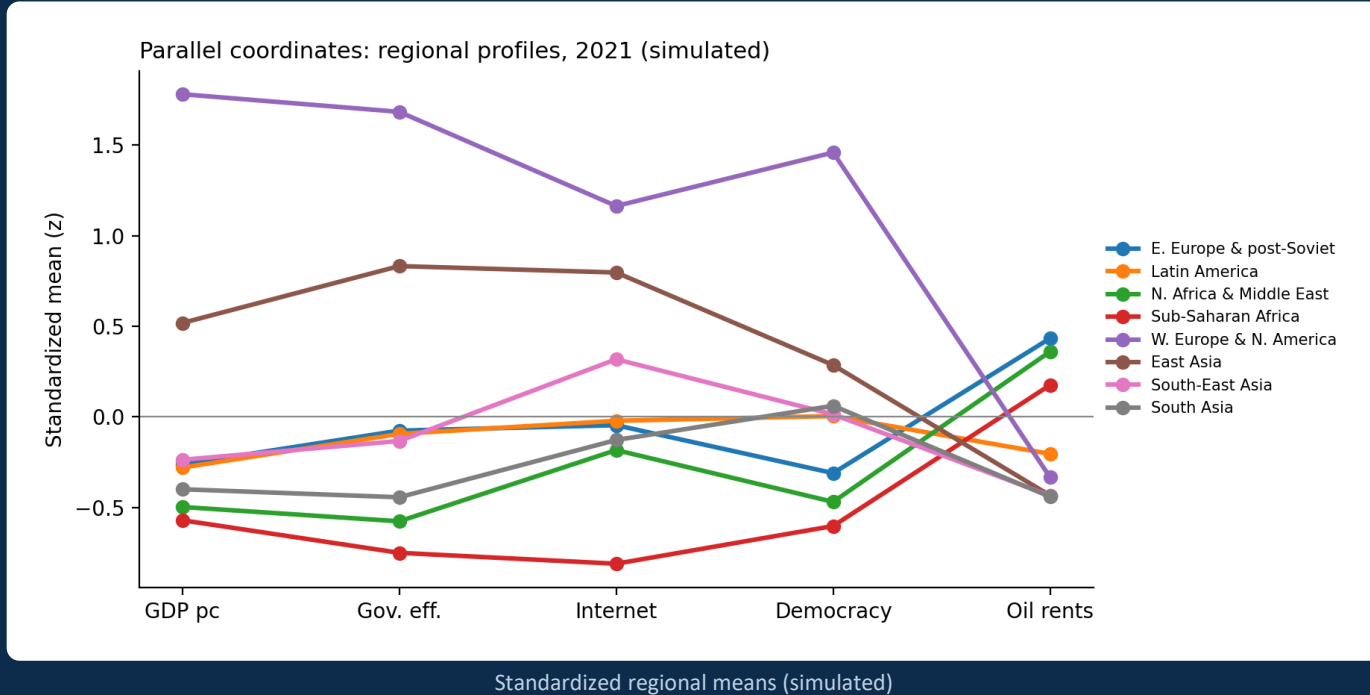
Small multiples over time

Small multiples: GDP per capita by country (gray) and regional median (color)

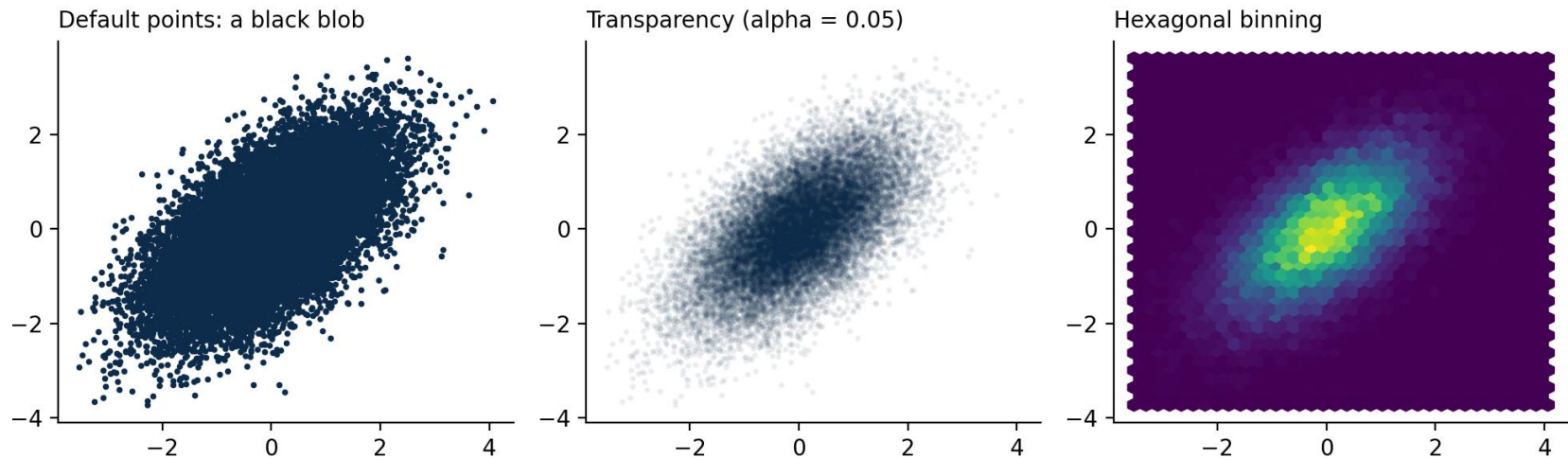


facet_wrap(~ region): countries in gray, regional median in color

Parallel coordinates



Too many points: overplotting



20,000 points: default, transparency, hexagonal binning

Graphs you can rotate, hover, and filter

plotly (R)

```
plotly(df, x = ~X15, y = ~X9, z = ~Y1, color = ~X18, type = "scatter3d", mode = "markers")
```

ggplotly

```
plotly::ggplotly(my_ggplot) # turns any ggplot into an interactive graph
```

gganimate

```
transition_time(year): animated bubble charts over time
```

Shiny

Web apps with sliders and menus for exploring data

Maps

```
sf + ggplot2 or leaflet: choropleth maps of country data
```

Principles of effective graphs

Do

- Show the data; label axes and units.
- Encode key comparisons by position and length (Cleveland & McGill, 1984).
- Start bar charts at zero.
- Use colorblind-safe palettes (viridis).
- Use log scales for skewed variables.

Avoid

- 3D bar and pie charts: they distort.
- Dual axes that exaggerate relationships.
- Truncated axes on bar charts.
- Chart junk: maximize the data-ink ratio (Tufte, 1983).
- Rainbow color scales for ordered values.

3D is valuable for surfaces and interactive exploration, not for decorating simple charts.

Match the color scale to the data

Sequential (viridis): ordered values



Diverging (red-blue): values around a center, e.g., correlations



Qualitative: unordered categories



- Sequential: ordered values.
- Diverging: values around a center (correlations).
- Qualitative: unordered categories (≤ 8).
- About 8% of men have some color vision deficiency.

ggplot2 recipes for GOV1

Histogram

```
ggplot(gov, aes(Y1)) + geom_histogram(bins = 40)
```

Boxplot / violin

```
ggplot(gov, aes(x = "", y = Y1)) + geom_violin() + geom_boxplot(width = .1) + stat_summary(fun = mean, geom = "point", shape = 18, color = "gold", size = 4)
```

By region

```
ggplot(gov, aes(reorder(factor(X19), Y1, median), Y1)) + geom_boxplot() + coord_flip()
```

Scatter + smoother

```
ggplot(gov, aes(X15, Y1)) + geom_point(alpha = .3) + geom_smooth()
```

Heatmap

```
corrplot::corrplot(cor(gov[, vars]), method = "color", type = "lower", addCoef.col = "black")
```

Pairs / ridges

```
GGally::ggpairs(gov[, c("Y1", "X7", "X9", "X15")]); ggribes::geom_density_ridges()
```

Key takeaways

- 1 Always plot first: summaries can hide everything (Anscombe).
- 2 Histograms, boxplots, violins, ECDFs, and QQ plots show distributions; Y1 is strongly right-skewed.
- 3 Heatmaps show linear pairwise correlations; follow up with scatterplots and scatterplot matrices.
- 4 Bubble charts, 3D views, ridgelines, small multiples, and parallel coordinates show several variables at once.
- 5 Encode by position and length, choose colors deliberately, and use 3D and interactivity where they add insight.

MODULE 2

Data Visualization

Roumen Vesselinov · JHBigData.com