Financial Econometrics II – Cross Section and Panel Data

Miscellaneous Advice & Thoughts on Empirical Work

Andreas Fuster
Swiss Finance Institute @ EPFL

SFI Léman PhD program – 2025

Things that are important but not in the usual econometrics textbooks

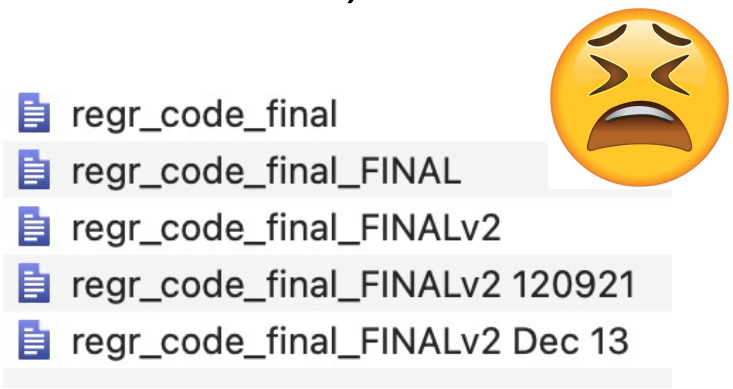
s:fi

1. Building good habits when doing empirical work
2. Generative AI (ChatGPT etc.)
3. Data visualization
4. (more to be added in future lectures)

Getting organized

s:fi

- Now is the time in your research career to build good workflow habits (it's hard to change later, trust me)
  - coding style (write your "convention")
  - documentation
  - folder organization & file naming
  - version control (e.g. Git/Github)
  - task management



- See e.g. https://kevinbryanecon.com/techstack.html (highly recommended!), https://aeturrell.github.io/coding-for-economists/wrkflow-rap.html, and https://shapiro.scholars.harvard.edu/file_url/174 (older)

- (This applies to all research that is not done with paper and pencil – but seems particularly important for the typical "empirical corporate finance" project.)

This has become even more important in recent years

- "Larger" projects, more code, more co-authors, …
- Journal requirements – generally need to share code, sometimes data. Most top journals now have data editors that will make sure everything checks out, with sources.
  - "where did we get this data set on local tertiary education completion rates that we use as a control variable in Table A19?"
- You don't want this to happen to you:

https://doi.org/10.1111/jofi.12868

First published: 12 December 2019

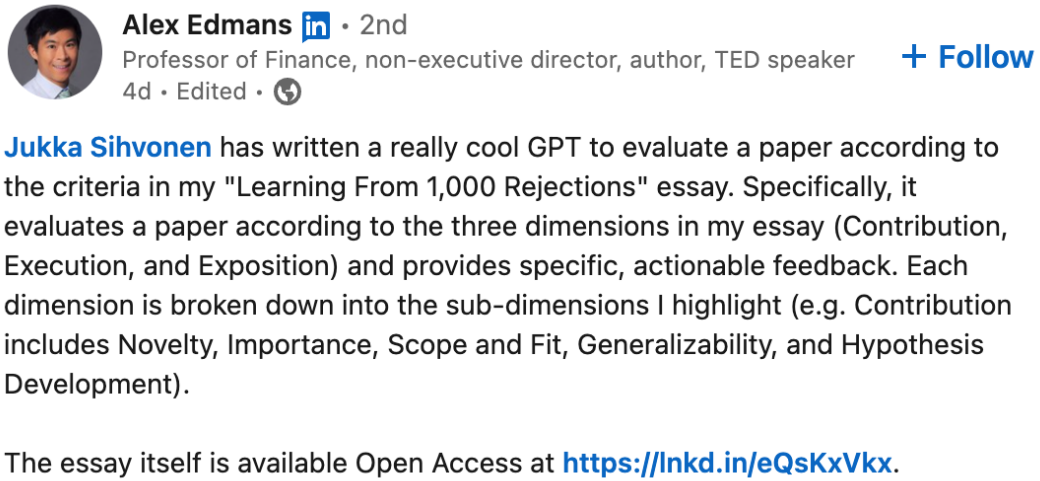
The authors hereby retract the above article, published in print in the April 2020 issue of *The Journal of Finance*. A replication study[1] finds that the replication code provided in the supplementary information section of the article does not reproduce some of the central findings reported in the article. Upon reexamination of the work, the authors confirmed that the replication code does not fully reproduce the published results and were unable to provide revised code that does. Therefore, the authors conclude that the published results are not reliable and that the responsible course of action is to retract the article and return the *Brattle Group Distinguished Paper Prize* that the article received. The authors deeply regret the damage this caused to the journal and the scholarly community. The specific

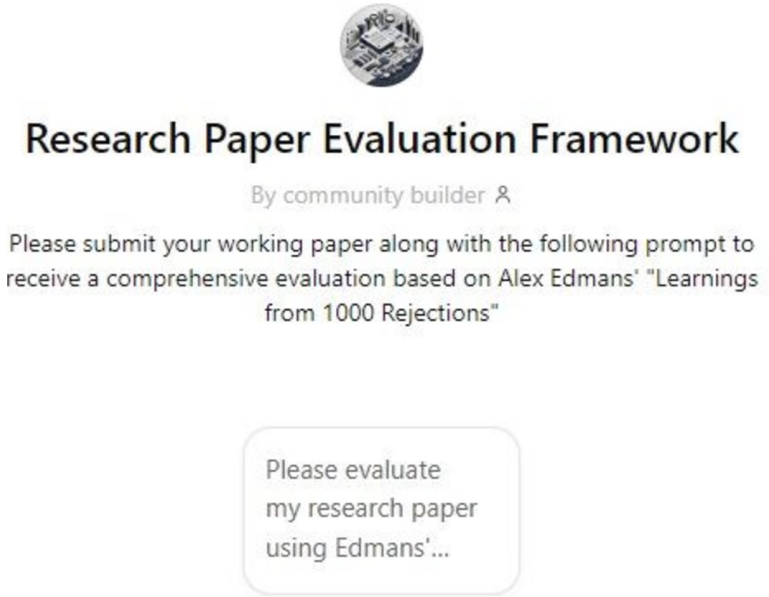
Generative AI (ChatGPT etc.)

- Generative AI tools will have a very large impact on how we do research in the coming years
- As new PhD students, you should become very familiar with these tools from the beginning of your research career!
- Already now, ChatGPT, Claude, etc. are quite impressive at tasks like:
  - coding (better in Python and R than in Stata)
  - writing
  - brainstorming
  - summarizing
  - and has become much better at literature reviews and math
- See e.g. Korinek (JEL 2023), “Generative AI for Economic Research” (incl. 2024 and 2025 updates, https://genaiforecon.org/)

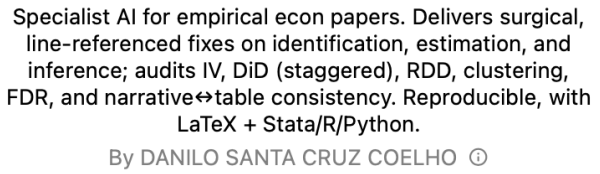
Generative AI (ChatGPT etc.)

- Example use case: have your drafts evaluated!





- Available on the GPT Marketplace: https://chatgpt.com/gpts
  - I've tried on one of my papers, it's useful
- Also useful: "Proof Patrol for Empirical Econ" →



- Paid tool: refine.ink
- In the future (or already now?) everybody will use such AI to check papers

Data visualization – binscatter / binsreg

s:fi

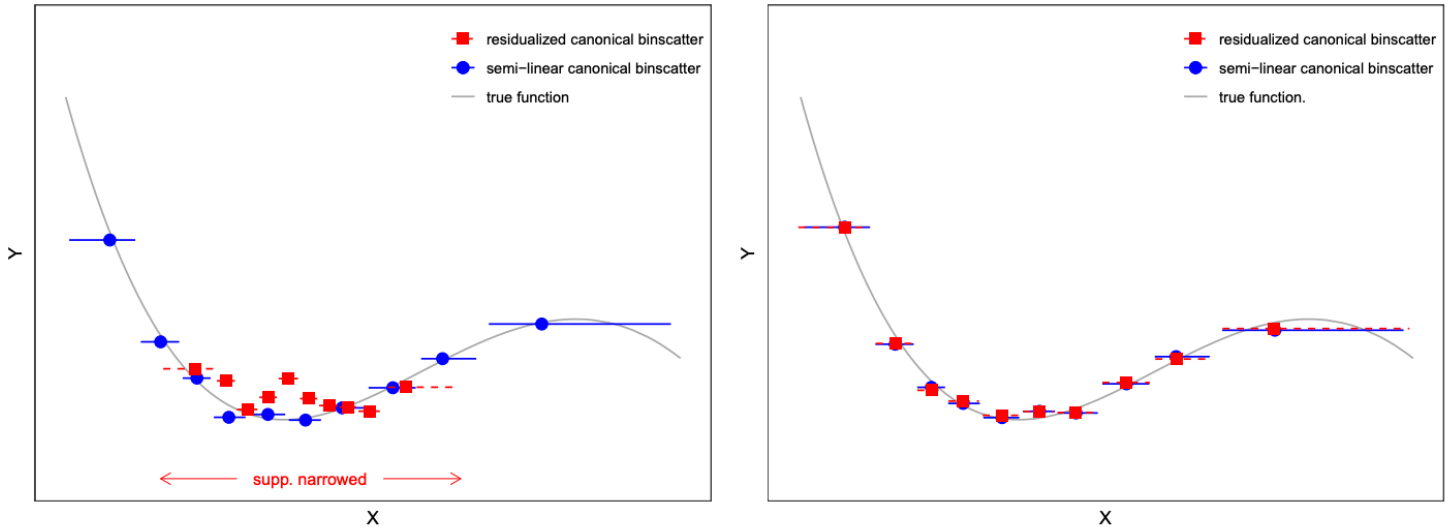
- Two purposes of good data visualization
  - understanding patterns in your data (exploration)
  - communication of research results
- Generally more challenging with large data sets
  - scatter plots with 1M observations not informative; overlaying local smoothers (e.g. lowess in Stata) takes forever
  - also, may want to do this after controlling for covariates / fixed effects, and sometimes also for binary dependent variables
- Approach that has become very popular: "**binscatter**"
  - sort *x* variable into bins (e.g. 20 or 50); show average *y* per bin
  - and using Frisch-Waugh-Lovell to "partial out" other covariates; but original binscatter command does not do this correctly

Worth familiarizing yourself with binsreg

(https://nppackages.github.io/binsreg/)

- (See Cattaneo & al. slides – attached in the back)
- Original binscatter covariate adjustment (based on partialing out other variables) biases resulting plot toward linearity (if g(x) nonlin.)

Figure 3: Comparison of Covariate Adjustment Approaches.



(a) x_i correlated with $\mathbf{w}_i$ (b) x_i independent of $\mathbf{w}_i$

Notes. This figure compares two approaches to covariate adjustment for binned scatter plots: semi-linear covariate adjustment and residualized covariate adjustment. Plot (a) compares the two approaches when there is non-zero correlation between x_i and the other covariates, $\mathbf{w}_i$, and demonstrates the biases introduced by naive residualization. Plot (b) compares the two approaches when x_i and the other covariates, $\mathbf{w}_i$, are independent. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The sample size is $n = 1,000$.

s:fi

- For more, see also these slides by Paul Goldsmith-Pinkham: https://github.com/paulgp/applied-methods-phd/blob/main/lectures/06_regression_2.pdf
  - Slides 28-35 have more general thoughts on effective visualizations in econ/finance papers
  - His design principles, which I generally agree with:

    1. Minimize tables
    2. Have describable goals for every exhibit
    3. Focus the reader and craft not-ugly figures
       - Ideally beautiful, but at minimum not ugly
    4. Do not *mislead* your readers

  - Rest of his PhD course slides are a great resource too

On Binscatter

Matias D. Cattaneo[1], Richard K. Crump[2], Max H. Farrell[3] and Yingjie Feng[4]

September 2021

(published in AER, 2024)

[1]Princeton University

[2]Federal Reserve Bank of New York. The views expressed here are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

[3]University of Chicago.

[4]Tsinghua University.

Introduction

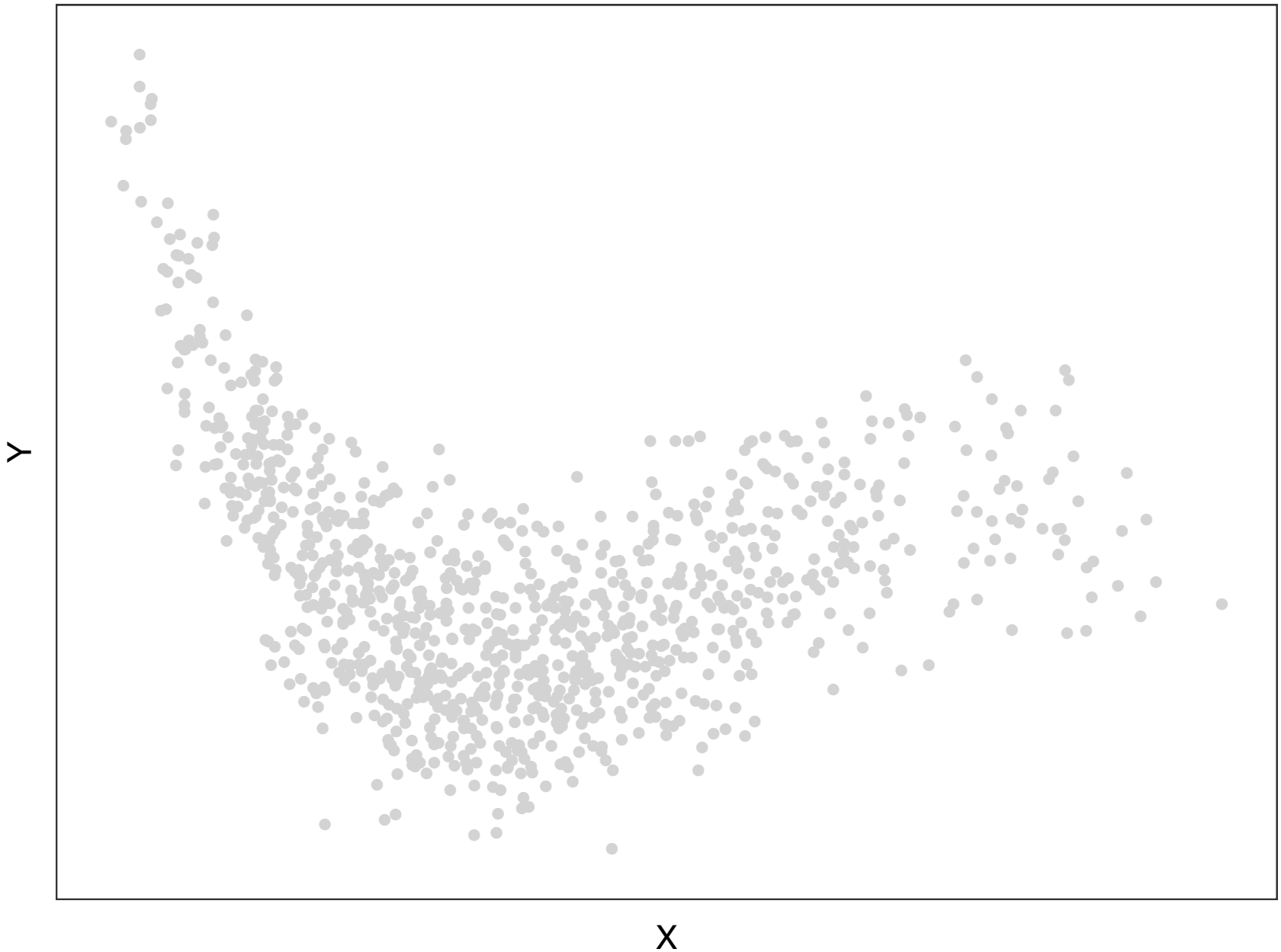
Binscatter is widely used in economics and other disciplines.

- Popularized by Chetty, Friedman, Rockoff, Saez, many others.
- Previous incarnations:
  - *Regressogram* (Tukey, 1961).
  - *Subclassification* (Cochran, 1968).
  - *Portfolio Sorting* (Fama, 1976).
  - *Regression Trees* (Friedman, 1977).
  - you tell me...
- Today: foundational, thorough study of Binscatter.
  - *Methodology*: guidance on valid and invalid current practices, and more.
  - *Theory*: novel strong approximation approach, and more.
  - *Practice*: new `Python`, `R` and `Stata` software (Binsreg package):

`https://nppackages.github.io/binsreg/`

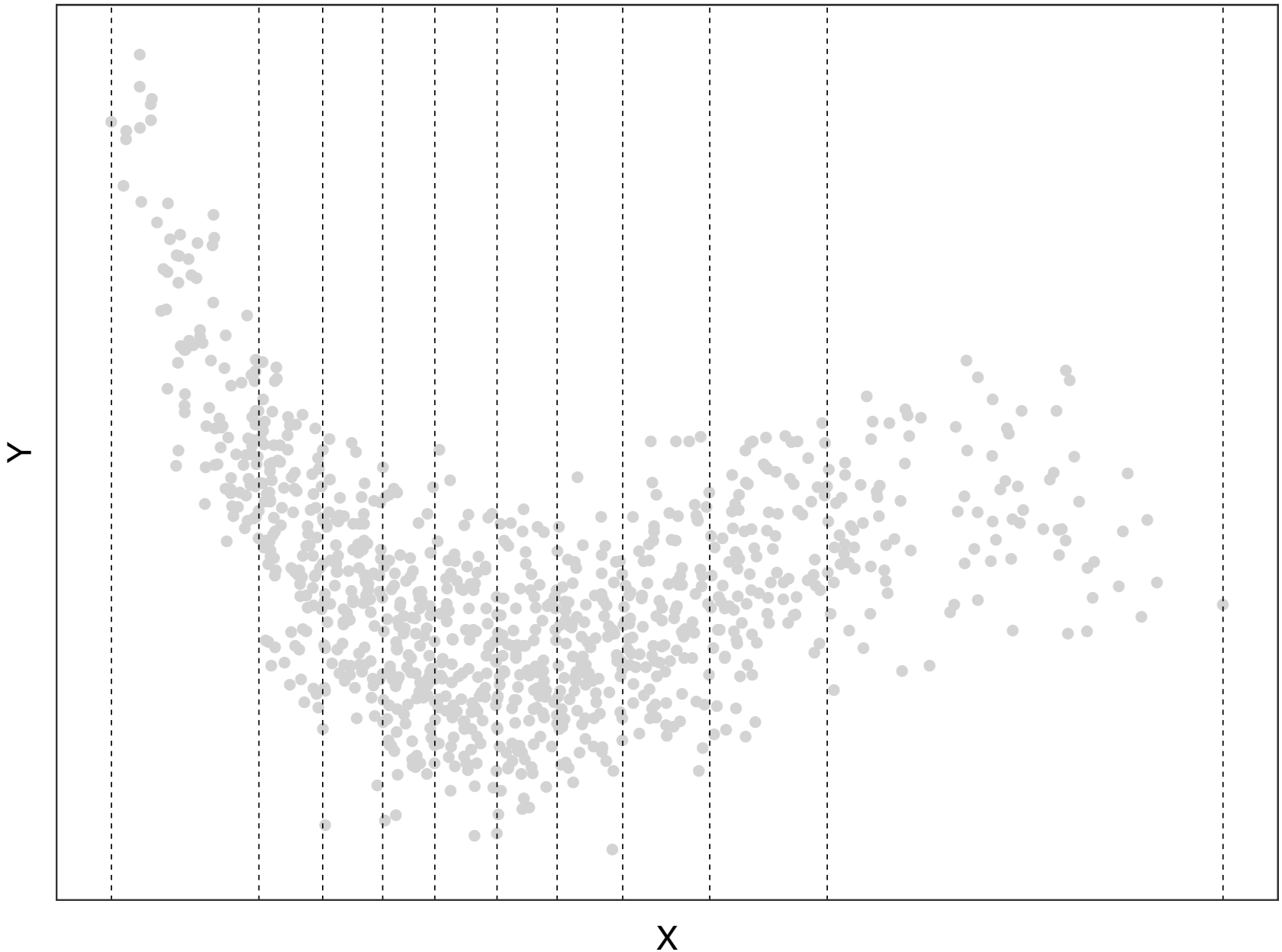
What is a binned scatter plot?

Step 1: Start with a familiar scatter plot



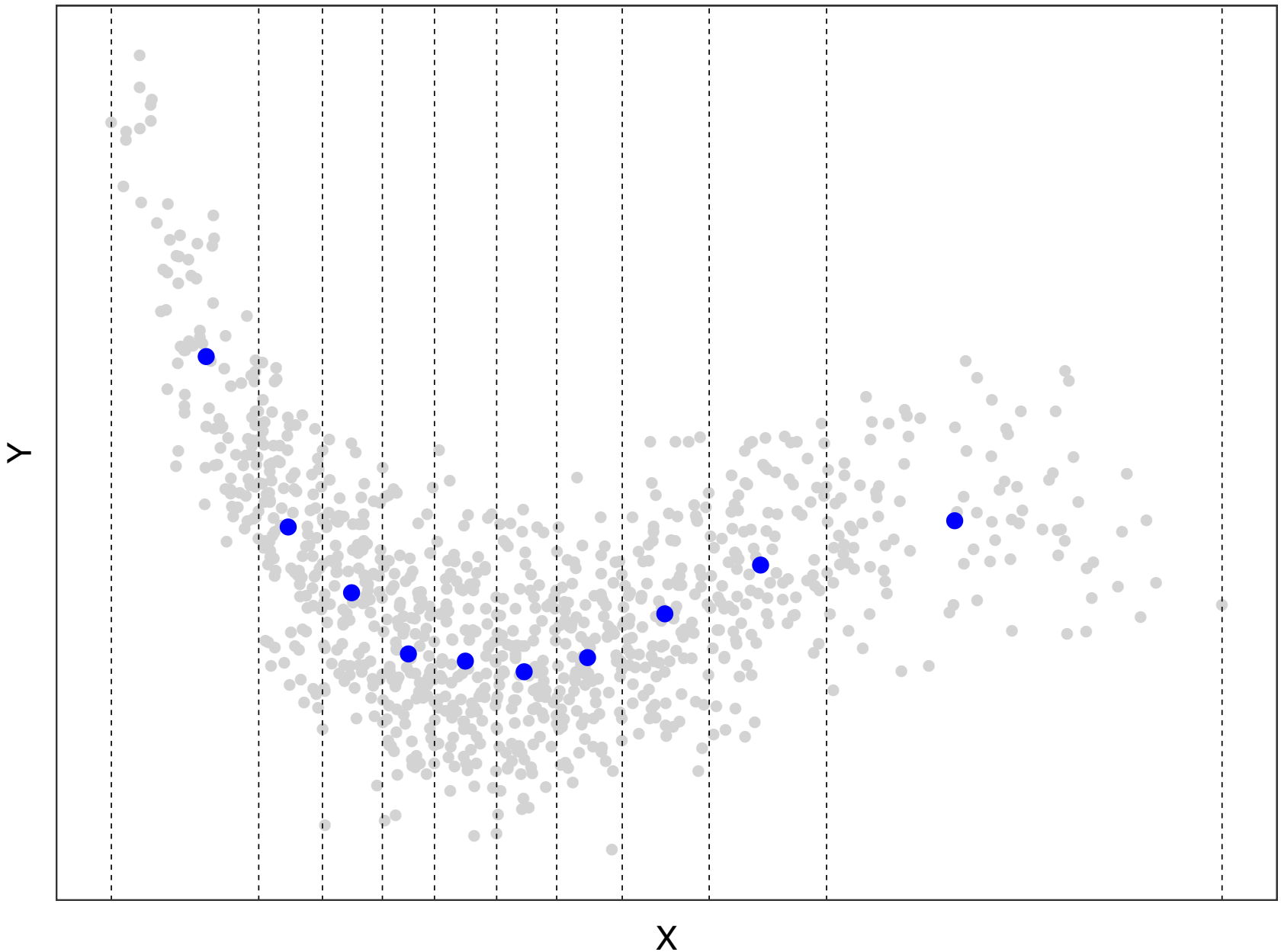
What is a binned scatter plot?

Step 2: Partition the support of X into bins



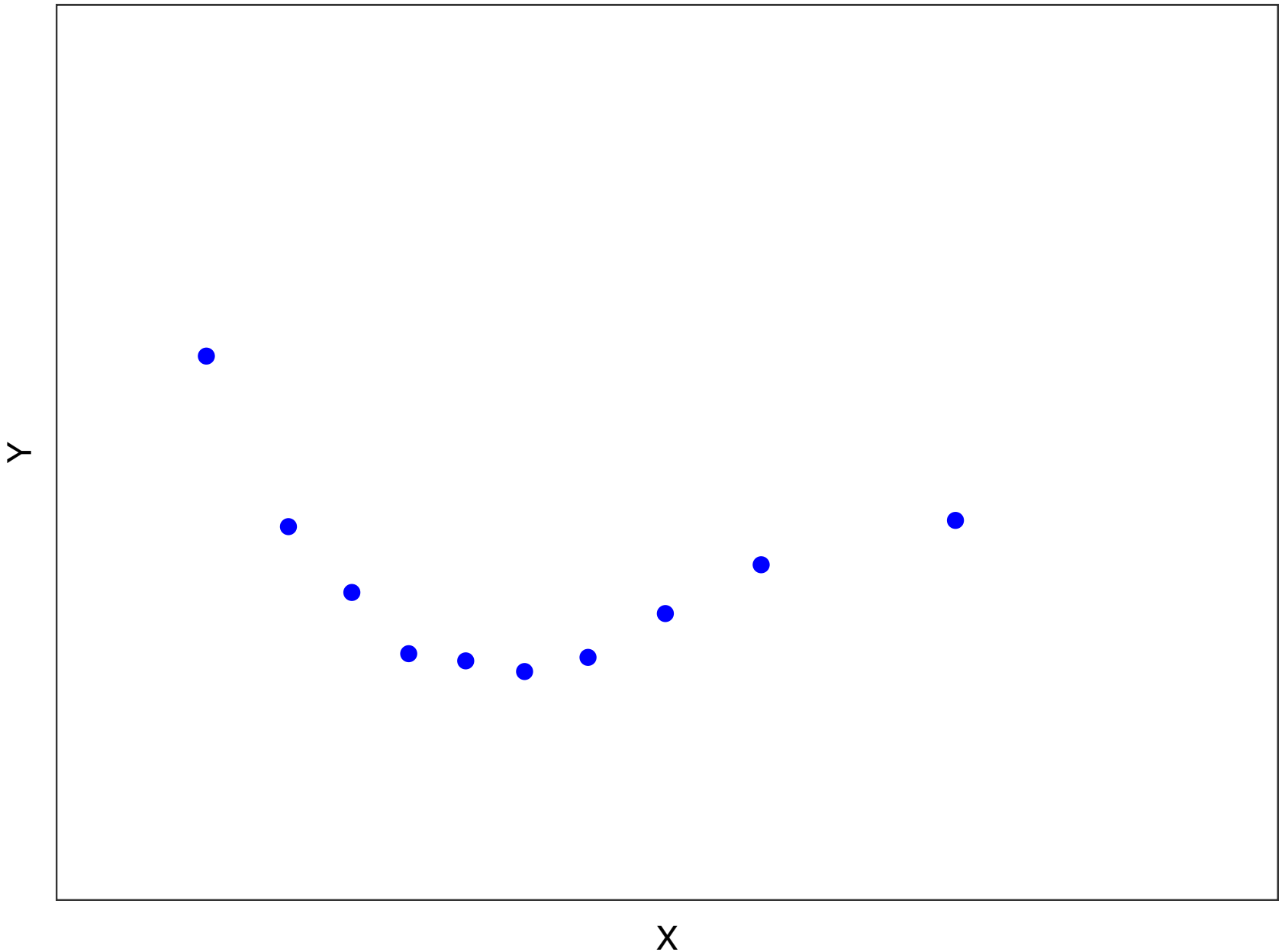
What is a binned scatter plot?

Step 3: Find the average Y in each bin



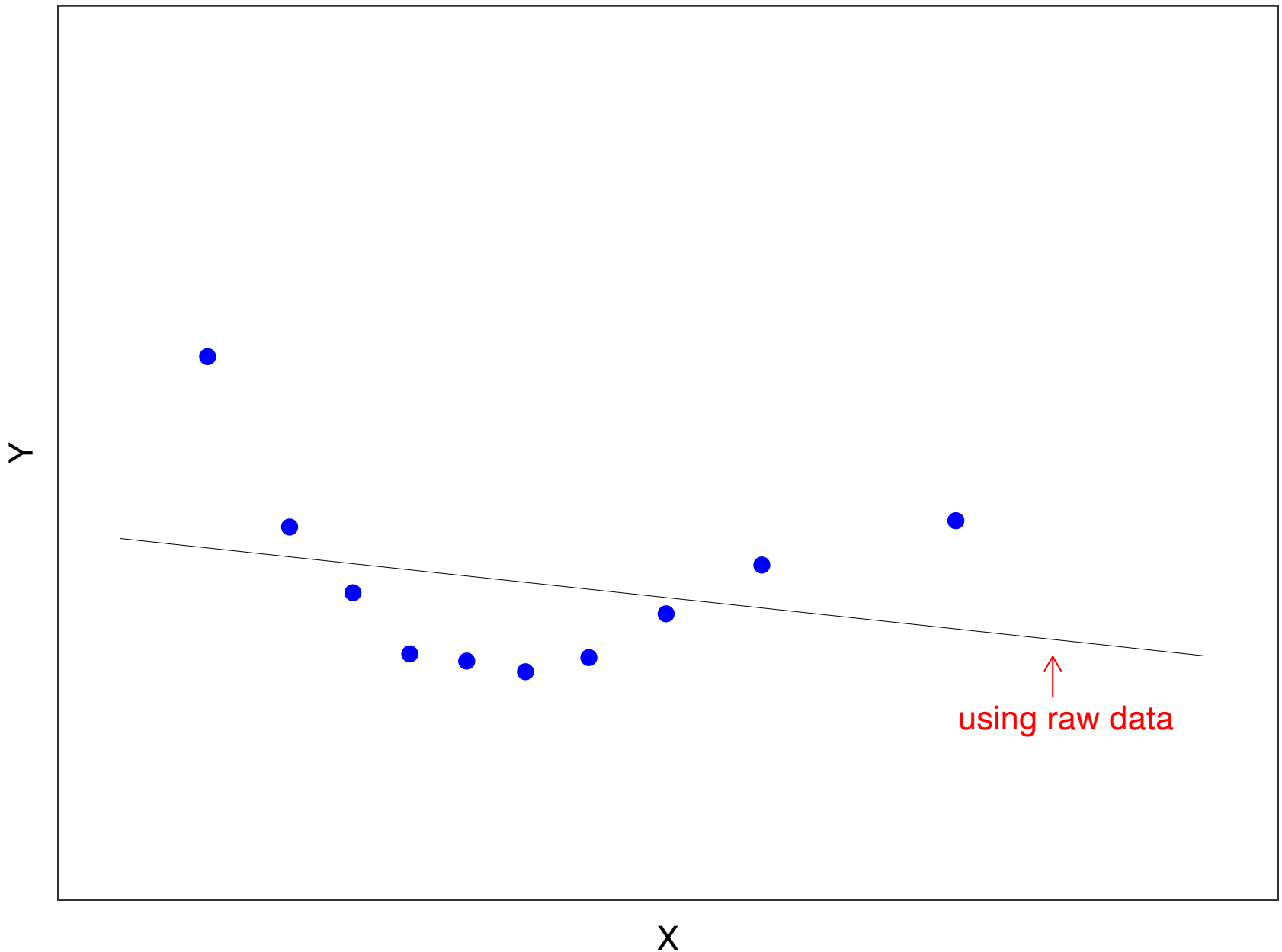
What is a binned scatter plot?

Step 4: Plot only bin means

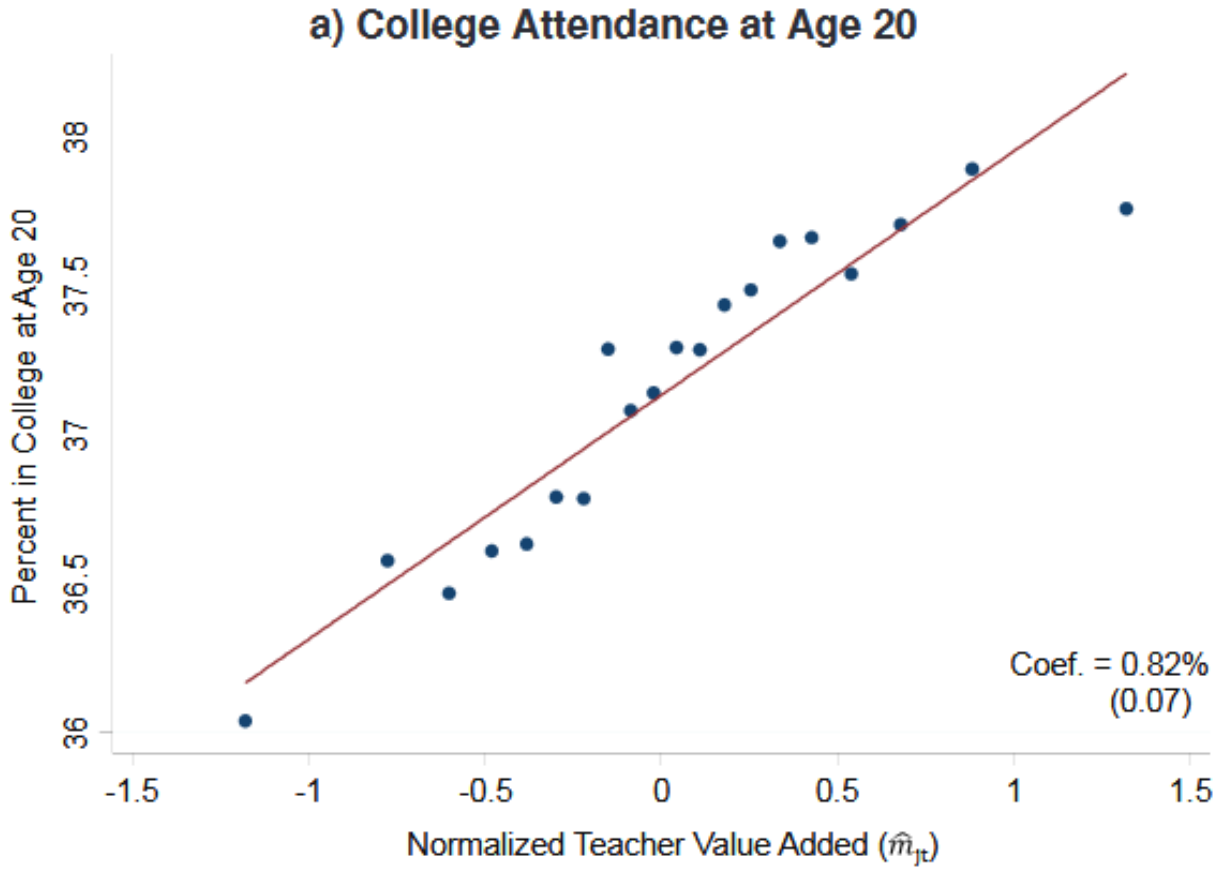


What is a binned scatter plot?

Step 5: Add a polynomial fit to raw data



Typical Example: Chetty, Friedman and Rockoff (2014, AER)



Note: $n = 4,170,905$ with # of bins $J = 20$

Overview: Contributions

1. Set up formal, general framework for studying **Binscatter**.
   - *Respects practice*: quantile-spaced binning, covariate adjustment.
   - *Extensions*: higher-order polynomial, smoothness-restricted approximations.
   - *Generalizations*: semi-linear QMLE (quantiles, logistic, etc.).
2. IMSE-Optimal choice of binning structure.
3. Valid point estimators, confidence intervals, and confidence bands.
4. Valid hypothesis testing of parametric specification and shape restrictions.
5. Novel theoretical results specifically developed for binscatter.
6. `Python`, `R`, and `Stata`software resolving valid and invalid current practices.